Simple Measurement Models for Complex Working-Memory Tasks

Klaus Oberauer

University of Zurich

Stephan Lewandowsky

University of Bristol and University of Western Australia

The research reported in this article was supported by funding from the University Research Priority Program "Dynamics of Healthy Aging" at the University of Zurich, as well as from the Swiss National Science Foundation (SNSF, grants number 100014_149193 and 100014_179002) to the first author, and a fellowship from the Royal Society to the second author. We are grateful to Danielle Pessach for assistance with collecting the data. Parts of the present research have been presented at the following conferences: Meeting of the Psychonomic Society, 2011 and 2017; Annual Interdisciplinary Summer Conference (ASIC) 2013; APS Conference 2016. An earlier version of the manuscript has been made available on PsyArXiv: osf.io/vkhmu.

Correspondence should be addressed to: Klaus Oberauer, University of Zurich, Department of Psychology – Cognitive Psychology; Binzmühlestrasse 14/22, 8050 Zürich, Switzerland, Email: k.oberauer@psychologie.uzh.ch

Abstract

We introduce a framework for simple measurement models for working memory, and apply it to complex-span and memory-updating tasks. Memory Measurement Models (M^3) use the frequency distribution across response categories to measure continuous memory strength along two dimensions: Memory for individual elements, potentially relying on persistent activation of unified representations, and memory for relations, relying on temporary bindings. Experiment 1 provides evidence for the validity of the parameters measuring these two dimensions of strength. The effects of experimental manipulations on these two dimensions can be captured by additional model parameters that reflect hypothetical processes affecting memory. Across five further experiments we illustrate how M^3 can be used to measure three such processes: The continued strengthening of memory representations during the retention interval (extended encoding), the dampening of encoding of irrelevant information (filtering), and the removal of irrelevant information from memory. In one experiment we compare young and old adults on complex-span tasks and working-memory updating. In both paradigms, old adults showed impaired memory for relations but no impairment in memory for individual elements. There was partial evidence for age differences in extended encoding and removal; there were no age differences in filtering. We suggest that M^3 offer a computationally efficient approach to identifying memory processes. All data and model codes are publicly available on the Open Science Framework: osf.io/vkhmu

Keywords: Working memory, measurement model, complex span, working-memory updating, aging

Simple Measurement Models for Complex Working-Memory Tasks

Usually the variables we measure in experiments do not directly reflect the constructs that we attempt to understand. For instance, we might be interested in people's ability to remember pictures, and therefore test picture memory through a recognition task. Measures of recognition accuracy, however, do not map one-to-one onto recognition ability, because other variables, such as response biases, also affect the observed variable. In general, our measurements are not process pure, or construct pure, but rather reflect a combination of influences from the latent process or construct of interest and other variables. Therefore, measurement models are needed to enable inferences from observed to latent variables. For instance, recognition researchers often rely on signal-detection theory as a measurement model for assessing a person's recognition ability (in a given experimental condition) from their behavior in a recognition task. Signal-detection theory provides a way in which response bias can be disentangled from recognition ability. Other examples of measurement models frequently used in cognitive psychology are sequential-sampling models of response times (S. D. Brown & Heathcote, 2008; Ratcliff & Tuerlinckx, 2002; Wagenmakers, van der Maas, & Grasman, 2007), and the process-dissociation procedure (Jacoby, 1991), which is a special case of the more general family of multinomial processing-tree models (Batchelder & Riefer, 1999; Buchner, Erdfelder, & Vaterrodt-Plünnecke, 1995).

All measurement models are deliberate simplifications, and rest on debatable assumptions, and they therefore all have well-known limitations. Yet, when we want to measure a variable of theoretical interest, there is no alternative to using a measurement model – rejecting use of a model is tantamount to implicitly using a naïve measurement model; that is, relying on the tacit assumption that the observed variable directly reflects the latent variable of interest. This naïve "model" is nearly always inappropriate, as just illustrated with the recognition example. Thus, notwithstanding their limitations, explicit

measurement models have the advantage of making explicit their assumptions about the mapping between latent and observed variables. Those assumptions are rarely outright wrong (as can be the case with implicit measurement models, cf. Loftus, Oberg, & Dillon, 2004) and are usually well justified by theoretical arguments and empirical tests.

Measurement models differ from explanatory models in that their main purpose is to measure latent variables in different experimental conditions rather than to explain differences between conditions (Oberauer & Lin, 2017). Measurement models usually allow most or all parameters to vary freely between experimental conditions with the aim of determining how the experimental manipulation affects the model parameters, which then permits inference to the effect on the latent variables of interest. In contrast, explanatory models are applied to all conditions together, aiming to explain the experimental effects with a single set of parameter values.[1] In comparison to explanatory models, measurement models are based on relatively sparse and fairly generic – yet testable – assumptions, they are applicable to a large range of phenomena, and are easy to use. A useful measurement model should be easy to fit to data for estimating parameters – to that end, an analytical expression is desirable. In some cases, as with simple applications of signal-detection theory, the parameters can be directly calculated from the data, thereby obviating the need for model fitting altogether. The purpose of this article is to introduce a framework for constructing measurement models for experimental paradigms used to study working memory, such as the complex-span paradigm (Daneman & Carpenter, 1980). We will refer to this framework as the M^3 framework (short for Memory Measurement Models).

The need for a measurement model for common working-memory paradigms arises because we use these paradigms for quantitative measurements of various aspects of working

[1] The contrast between measurement models and explanatory models should be understood as a continuum: Models in which some parameters are constrained to be the same across conditions, whereas others are free to vary, constitute intermediate cases.

memory. In individual-differences studies, complex span and other tasks are used to measure a person's working-memory capacity. In experimental studies, these tasks are used to measure specific mechanisms or processes assumed to play a role in working memory, such as memory for order, or resistance to distraction. In all these applications we must make an inference from the manifest variables we observe to the latent variables we are interested in. Without an explicit measurement model, we cannot do better than to take the manifest variable as a proxy for the latent variable, at best accompanied by an acknowledgement that our measurement is far from process-pure (Conway, Kane, & Engle, 2003). An explicit measurement model is needed to get closer to a process-pure assessment of latent variables of theoretical interest. Whereas measurement models have recently been developed for some working-memory tasks (Cowan, Blume, & Saults, 2013; Oberauer, Stoneking, Wabersich, & Lin, 2017; Zhang & Luck, 2008), they are still lacking for the most commonly used paradigms. The M^3 framework fills that gap.

Although typically relying on few assumptions, measurement models are not theoretically neutral, and often their core assumptions are a matter of intense controversy. Signal-detection theory applied to recognition, for instance, relies on the assumption that recognition decisions are based on a memory signal that varies in strength continuously, such that the person must set a criterion to arrive at a binary old-new decision. One can reject this assumption and instead endorse a discrete-state model of memory, as incorporated in high-threshold models of recognition (Bröder & Schütz, 2009). According to high-threshold models, a person either clearly remembers having experienced the recognition probe, or has no memory whatsoever, in which case they resort to guessing. There is an ongoing debate whether signal-detection based or discrete-state based theories of recognition are more appropriate (Dube & Rotello, 2012; Kellen & Klauer, 2014, 2015; Rouder et al., 2008; Wilken & Ma, 2004; Wixted, 2007).

More generally, discrete-state models of memory can be implemented in multinomial processing-tree models, because the latter rest on the assumption that cognitive processes traverse probabilistically through a series of discrete states, and the final state determines the decision for an overt response (e.g., saying “old” or “new”). Multinomial processing-tree models of memory have been applied not only to recognition but also to recall (e.g., Schweickert, 1993). Thus, there are well-developed measurement models for both recognition and recall based on the discrete-state assumption, and we have signal-detection models of recognition to represent the alternative continuous-strength assumption. However, to our knowledge so far there is no framework for building measurement models of recall on the assumption of continuously varying memory strength. Here we propose such a framework because the implications of continuously varying strength are far from trivial.

We need this class of measurement models because the large majority of more detailed explanatory models of recall are built on the assumption of continuously varying memory strength, both in the field of working memory (Burgess & Hitch, 1999, 2006; Farrell & Lewandowsky, 2002; Henson, 1998; Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012; Page & Norris, 1998) and in research on episodic memory (G. D. A. Brown, Neath, & Chater, 2007; Farrell, 2012; Hintzman, 1986; Raaijmakers & Shiffrin, 1981; Sederberg, Howard, & Kahana, 2008; Shiffrin & Steyvers, 1997). In particular, these models share the assumption that multiple recall candidates are activated to different degrees by the available retrieval cues and compete for being chosen for recall according to their degree of activation. In the serial-recall literature, this competition is often referred to as *competitive queuing* (Houghton, 1990). This competitive selection mechanism can only be captured by a model in which representations differ in continuously varying strength of activation at test. Competitive selection is a core mechanism in M^3.

Our goal is to develop the M^3 framework for the analysis of working-memory tasks, and we will demonstrate its application to experiments with two relevant paradigms, complex span and working-memory updating. In principle, the M^3 framework could also be extended to paradigms for studying recall from episodic long-term memory, such as delayed free recall or probed recall.

The Basic Model

The M^3 framework rests on two generic assumptions about recall from working memory. First, we think of each recall attempt as the selection of a response from a set of candidates. Sometimes the set of candidates is implied by the material—for instance, when the task is to recall a list of digits in order, the digits 1 to 9 (or sometimes 0 to 9) form the candidate set. In other cases, the candidate set is constructed by the person. For instance, when asked to recall a list of words, a person's entire vocabulary is in principle eligible for the candidate set. The candidate set may however be narrowed down if the person notices, for instance, that the words were all concrete single-syllable nouns.

The second assumption is that people select from the candidate set according to the relative activation of each candidate representation at test. The activation level of each candidate is a continuous variable reflecting the strength of evidence from memory in favor of selecting a candidate. Hence, activation is similar to the signal strength in favor of an "old" response in signal-detection models of recognition, except that in M^3 all potential candidates have their own distinct signal strength.

For the basic model of recall from working memory we consider two sources of activation, based on two kinds of information: memory for individual elements and memory

for relations.[2] By memory for elements we mean information about which individuated events have been experienced during the episode relevant for recall, regardless of any relations of that event to other events or to context. In a typical experiment, memory for elements means information about which list items have been presented in the current trial. In general, an individuated event is any unit in the stimulus material for which the person has a unified representation in long-term memory (i.e., a chunk in the sense of Miller, 1956). Memory for relations, by contrast, refers to information about how an event relates to other events and to its context. In a typical working-memory task this includes knowing the serial position of an item in a list, or the location of a visual object in space.

We can tentatively map memory for elements and for relations to different hypothetical mechanisms of retention in working memory. Many theories of working memory assume that short-term retention is accomplished by persistent activation of a selected set of representations (Curtis & D'Esposito, 2003; Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005; Wei, Wang, & Wang, 2012). Persistent activation is a natural candidate for maintaining memory for elements: By definition, an individual chunk has a unified representation that can be temporarily activated. In a model in which this activation can be sustained for some time after encoding, a representation's level of activation can be used to determine whether it has been used recently, but it contains no information about its context.[3]

[2] Other closely related terms are item memory vs. order memory (when the relation is one of order) (Marshuetz, 2005), and item vs. associative memory (Gronlund & Ratcliff, 1989). Here we use the term *element* instead of *item*, because we reserve the term item to denote an element that a person is asked to remember, as opposed to a *distractor*, which is an element the person is asked to process but not to remember.

[3] It is important to distinguish between the concept of *persistent activation* as a mechanism for maintaining the occurrence of an element in memory, and the concept of *activation at retrieval*, which reflects the total strength of evidence from memory for each retrieval candidate. Persistent activation of a representation contributes to its activation at retrieval, together with its re-activation through its binding to the currently used retrieval cues.

Another mechanism of retention in working memory is to establish temporary bindings or associations between representations of contents and some representation of context, such as bindings between list items and their list position. This mechanism is used in most contemporary models of memory for serial order, and has received strong empirical support (Farrell & Lewandowsky, 2004; Hurlstone, Hitch, & Baddeley, 2014; Lewandowsky & Farrell, 2008). Obviously, bindings are ideally suited to represent relations. We note that the mapping of memory for elements to persistent activation, and memory for relations to bindings, is only tentative because some form of relational memory can be accomplished by gradients of activation, such as a primacy gradient of activation across list items for representing their serial order (Page & Norris, 1998), although this capability is very limited (Oberauer & Lewandowsky, 2011). Conversely, memory for the occurrence of individual events could rest on bindings between these events and a representation of the global context of the relevant episode (e.g., a trial context distinguishing the current from previous trials).

To apply the M^3 model, we distinguish categories of possible responses, that is, categories of elements in the candidate set, for which the model predicts different levels of activation at test. As an example, consider the task of remembering a list of words in their correct order (i.e., a so-called simple-span task). At test, participants recall the list by selecting the word for each output position (i.e., each position in the recall sequence) from a set of candidates, including all list words, together with a number of new words, which we will refer to as not-presented lures (NPLs). We can distinguish three categories of responses: The *correct* item at any given position in the recalled list, *other items* from the current memory set, and *NPLs* that were not included in the current memory set. The basic model predicts the frequencies of responses in these three categories (in the serial-recall literature, these responses are referred to as *correct-in-position*, *transposition* error, and *extralist intrusion* error). According to the model, each retrieval candidate is supported by a combination of

sources of activation at retrieval, depending on which of these three categories they belong to, which can be expressed by the following equations:

$$\begin{aligned} A(correct) &= b + a + c, \\ A(otheritem) &= b + a \\ A(NPL) &= b. \end{aligned} \quad (1)$$

Here, *b* is the baseline activation assumed for all candidates. It is a fixed parameter that serves as a scaling parameter in the model; we set it arbitrarily to 0.1.[4] The remaining two terms are free parameters to be estimated from data. Parameter *a* reflects the strength of memory for individual elements, which in the case of serial recall is memory for which words have been presented in the current trial. Parameter *c* reflects the strength of memory relations, which in serial recall represents memory for which word has been in the currently to-be-recalled list position. As mentioned above, the most successful models of serial recall assume that list items are bound to a temporal context marking their position in the list. At recall, the context is re-instated in forward order, such that each position cues the item bound to it. In the model, the *c* parameter reflects the strength of evidence conveyed to an item when cued by the context to which it was bound; in serial recall this is arguably the item's temporal context.

The model incorporates the simplifying assumption that the context cue used at each output position is bound only to the correct item, and therefore *c* is only added to A(correct). This assumption is a simplification because the temporal-context cues are likely to overlap, such that each cue also cues neighboring memoranda (Burgess & Hitch, 1999), and in complex-span tasks the distractors are also bound to the positions of nearby memoranda (Oberauer, Farrell, Jarrold, Pasiecznik, & Greaves, 2012).

[4] This scaling parameter is analogous to the within-trial SD of drift rate (parameter *s*) in the diffusion model, which is fixed to set the scale for other parameters (Ratcliff & Rouder, 1998)

To translate the relative strength of activation at retrieval into probabilities for each response category, we use Luce's choice rule:

$$p(i) = \frac{A(i)}{\sum_{j=1}^{n} A(j)} \quad (2)$$

The sum in the denominator runs over the *n* recall candidates (i.e., the individual candidates in all response categories) because *A(j)* represents the activation value of candidate *j*, and the choice is between retrieval candidates. An important aspect of this model is that A is expressed on a ratio scale. That is, the zero point on that scale is meaningful, and it indicates the absence of any memory strength for the candidate in question. This desirable property eliminates other candidate choice rules from consideration (See Appendix A for details).

For application of the model, it is important to have a well-defined set of recall candidates so that it is known over which candidates the sum in the denominator of Luce's choice rule is to be taken. For some materials (e.g., digits, letters, spatial positions in a grid) there is a naturally limited candidate set. For others (e.g., words) the candidate set is potentially very large, up to all words in the language. In the experiments reported below, we asked participants to recall word lists by selecting the words from an array of candidates on the screen, which enabled us to define a more restricted candidate set. The basic model can be used to measure two latent variables: Parameter *c* reflects the strength of relational memory (e.g., the strength of binding between an item and its position), and *a* reflects the strength of memory for individual elements. The two memory-strength variables *c* and *a* are estimated relative to the scaling parameter *b* (baseline strength). The model can distinguish changes in the two strength parameters by the changes in the response proportions they imply: An increase in *c* implies an increase of correct responses relative to other-item responses and NPLs. In contrast, an increase in *a* implies an increase in other-item responses relative to NPLs.

Measurement models built within the M^3 framework can be used to gauge the effect of experimental manipulations on the two strength parameters. This involves incorporating additional parameters that modify the two strength parameters, thereby capturing the experimental effects on them. To illustrate how this works, consider a simple serial-recall experiment with an experimental manipulation of memory strength (e.g., varying the presentation duration per item). The top panels in Figure 1 show simulated data generated by three versions of the basic M^3, one in which the manipulation only increases the strength of elements (a = 0.5 vs. 0.8), one in which it increases only binding strength (c = 7 vs. 11) and one in which it increases both parameters. We fit these data with an M^3 that captures the experimental effects through two change parameters, Δa for the change in a between experimental conditions, and Δc for the change in c. The bottom panels of Figure 1 show the posteriors of estimates of these change parameters, which accurately recover the selective influence of the experimental manipulation on one parameter in the first two simulations, and on both parameters in the third. Note in particular how the visually quite small effects in the data translate into unambiguous and large differences in the parameter estimates (for details of the Bayesian implementation of M^3 see below).

Extended Measurement Models for Complex WM Tasks

More complex WM paradigms can yield richer data, which we can leverage to measure additional processes through M^3. Consider a complex-span task in which participants are asked to remember a list of words, and after presentation of each list word they engage in a distractor task that involves processing one or several other words (e.g., simply reading these words aloud, or making a judgment on them). At test, the person is asked to reproduce the memory list by selecting the list items from a set of candidates. The candidates include all list items, all or some of the distractors, and NPLs. We can now distinguish five categories of responses: At each output position a person could select the correct item (for the to-be-

recalled position), another item from the memory list (a transposition error), a distractor word from the to-be-recalled position, a distractor word from another position, or an NPL. We can now ask how the status of a word – as a memory item or as a distractor – affects the strength of memory for its occurrence as an element (parameter *a*), and the strength of its binding to its temporal context (parameter *c*).

By virtue of being processed, distractors are encoded into working memory, but probably not as strongly as memory items (Oberauer, Farrell, et al., 2012). Several theorists have proposed that there is a degree of control over which contents are admitted into working memory, and that individuals might differ in the efficiency of such a gating or filtering mechanism (Awh & Vogel, 2008; Hasher, Zacks, & May, 1999; Rac-Lubashevsky & Kessler, 2016). To capture the possibility that distractors are encoded with reduced strength relative to memory items, for distractors we multiply *c* and *a* with a filtering parameter *f* (assumed to have a value between 0 and 1) that reflects the proportional reduction of list memory by (partially) filtering the encoding of distractors. In the extreme case that distractors are not encoded into working memory at all, *f* would be estimated to zero. The extended model equations are:

$$\begin{aligned} A(correct) &= b + a + c, \\ A(other\ item) &= b + a \\ A(distractor\ in\ position) &= b + f(a + c) \\ A(other\ distractor) &= b + fa \\ A(NPL) &= b. \end{aligned} \tag{3}$$

The distinction between *distractors in position* and *other distractors* is motivated by the assumption that distractors, to the extent that they are encoded into working memory at all, are also bound to the temporal context of the corresponding list item. Support for this assumption comes from the observation of a *locality constraint* on distractor intrusions: When

a distractor is erroneously recalled instead of an item, the correct item is more likely to be replaced by a distractor close to it in the input sequence (i.e., the sequence of events at encoding) than by a distractor further removed. Sometimes the most prevalent distractor intrusions come from the distractors immediately following the item they replace (Oberauer, Farrell, et al., 2012). In other instances, distractor intrusions come predominantly from distractors immediately preceding and those immediately following the replaced item (Oberauer & Lewandowsky, 2016). The latter pattern was observed for the three complex-span experiments analyzed here. Therefore, we categorized responses choosing distractors immediately preceding or following the items they replace as *distractors in position*, and all other distractor responses as *other distractors*.

In addition to the theoretical assumptions about memory that we built into the M^3 framework, we also had to make a few auxiliary assumptions necessary to make the models work. In Appendix A we explain and justify these assumptions.

Bayesian Hierarchical Modelling

We implemented all M^3 variants as Bayesian hierarchical models (Lee & Wagenmakers, 2014). In a hierarchical model, parameters of individuals are not estimated independently, but rather are modeled as samples from a distribution that is specified by a mean and a dispersion parameter. The mean and dispersion of that distribution are so-called population-level parameters (a.k.a. hyper-parameters). In this way, the model allows for individual differences in parameter values while constraining them to belong to a common distribution. At the same time, we obtain parameter estimates on the group level (e.g., the mean of the distribution from which individual parameters are sampled) that we can interrogate for experimental effects or group differences.

Applying the models with Bayesian estimation methods has several advantages over Maximum-Likelihood fitting methods. First, rather than point estimates of parameters, we obtain posterior probability distributions on parameter values, telling us how probable each possible parameter value is in light of the data (see Figure 1). Second, implementing hierarchical models is particularly easy in a Bayesian framework, because drawing parameters from (prior) distributions lies at the heart of Bayesian modeling. In a hierarchical model we simply treat the population-level parameters as priors of the individual-level parameters. Third, Bayesian modeling uses very efficient algorithms – so-called Markov-Chain Monte-Carlo (MCMC) samplers – for searching high-dimensional parameter spaces, so that hierarchical models, which often have a large number of parameters, can be estimated with little difficulty.

We applied the models to the data using JAGS 4.2 (Plummer, 2016) together with the R2jags package (Su, 2015) in R (R_Core_Team, 2017). For each model we ran 3 MCMC chains, each with 30,000 sampling steps (after 5,000 burn-in steps that are discarded). Convergence of the chains for the population-level parameters was checked by calculating $\hat{R}$ (Gelman & Rubin, 1992) using the *gelman.diag* and *gelman.plot* functions in the CODA package (Plummer, Best, Cowles, & Vines, 2006); good convergence is indicated by $\hat{R}$ not exceeding 1 by much. Some models for which convergence was not satisfactory ($\hat{R} > 1.05$) were run again with 100,000 chains. After that, large majority of models, and all models for which we report parameter estimates below, had $\hat{R} < 1.05$. Model comparison was based on the WAIC information criterion (Watanabe, 2010), which is suited for hierarchical models and has better statistical properties than the more frequently used DIC (Gelman, Hwang, & Vehtari, 2014). Smaller WAIC values reflect better model fit. To facilitate interpretation of WAIC differences between models we ran model comparison analyses, reported at the end of

this article, which showed that WAIC differences > 10 can be interpreted as strong evidence in favor of the winning model.

Overview of Experiments

Our first experimental test of the basic model defined in Equations 1 and 2 involved an analysis of the responsiveness of its parameters to selective experimental manipulation. If these parameters measure the latent variables they are meant to measure, then experimental manipulations that we expect to influence only one latent variable should selectively affect only the parameter thought to measure that variable (Batchelder & Alexander, 2013; Heathcote, Brown, & Wagenmakers, 2015). Experiment 1 carries out such a test of selective influence for parameters *a* and *c*.

The extensions to the basic model depend on the experimental paradigm and design, and on what possible effects one considers the experimental manipulations to have on the memory strength parameters. Experiments 2 to 5 illustrate how M^3 tailored to specific experiments can be used to test theoretical assumptions about the effects of experimental manipulations on the two core parameters, *a* and *c*. Experiments 2 to 4 serve to analyze effects on performance in the complex-span paradigm; Experiment 5 provides data for testing a measurement model for another experimental paradigm often used to study working memory, viz. the memory-updating paradigm (Kessler & Meiran, 2008; Kessler & Oberauer, 2014; Oberauer, 2003; Salthouse, Babcock, & Shaw, 1991).

Measurement models can also be used to investigate individual differences in the theoretical constructs represented by its parameters. For instance, we can ask how individuals differ in the strength of memory for elements and of memory for relations, or in their ability to filter distractors. The final experiment reported here (Experiment 6) illustrates this use of

measurement models of complex-span and memory-updating tasks for investigating age differences in working memory.

Experiment 1: Selective Influence on Parameters

Design

Experiment 1 used a complex-span task: Participants remembered lists of five words. Each memory word was followed by a distractor word that participants had to process (i.e., read aloud) but were instructed not to remember. In the control condition the distractors were new words. In the *old-reordered* condition, the distractors were the same words as the memory items, but in a different order. In the *old-same* condition, the distractors were the same as the memory items, presented in the same order, so that each memory word was followed by a distractor identical to it. At test, participants selected the five memory words in order from a candidate set of 15 words. In the control condition, the candidate set consisted of the five memory words, the five distractors, and five new words (NPLs); in the old-reordered and old-same condition, in which no new distractors were used, the candidate set consisted of the five memory words and 10 NPLs.

On the assumption that distractors are encoded into working memory, the old-reordered condition should increase the strength of each memory item as an element, but not their bindings to their positional context. Therefore, compared to the control condition, the old-reordered condition should selectively increase parameter a. The old-same condition should increase the strength of bindings between each memory item and its position, because that same binding is re-encoded when processing the distractor. Hence, this condition should increase parameter c relative to the control condition. In addition, it arguably should also increase memory for elements (parameter a). Details of the methods of Experiment 1 can be found in Appendix B.

Measurement Models

We applied the basic measurement model to predict the proportion of responses in three categories: The correct word, other list items, and NPLs. For the control condition only, selection of a distractor formed a fourth response category; for the other two conditions, in which the distractors were repetitions of the memory items, this response category could not be defined. Because here we were not interested in distractor selections, we simply augmented the basic model by a parameter *d* to represent the larger activation of distractors relative to NPLs in the control condition:

$$\begin{aligned} A_j(correct) &= b + (a + \Delta a_j) + (c + \Delta c_j), \\ A_j(other\ item) &= b + (a + \Delta a_j) \\ A_j(NPL) &= b \\ A_0(distractor) &= b + d, \end{aligned} \tag{4}$$

where *j* is an index for the condition (*j* = 0 for *control*, *j* = 1 for *old-reordered*, *j* = 2 for *old-same*). The potential differences in *a* and *c* between the two *old* conditions and the control condition were captured by the Δa and Δc parameters.

In a first pass, we freely estimated Δa_1 and Δc_1 for the old-reordered condition, and Δa_2 and Δc_2 for the old-same condition, so that *a* and *c* could assume different values in all three conditions. We compared this model to a version in which we constrained the parameters to behave as expected on the assumption of selective influence: $\Delta c_1 = 0$ for the assumption that old-reordered only increases *a*, and $\Delta a_2 = \Delta a_1$ for the (more tentative) assumption that both old conditions increase *a* equally.

Results and Discussion

The top panel of Figure 2 shows the proportions of responses in each category for the three conditions. The bottom panel presents the posterior distributions of parameter values

from the unconstrained M^3. Parameter *c* behaved exactly as predicted: It increased relative to the control condition in the old-same condition, as reflected in the positive Δc_2 (red horizontal bar for 95% highest-density interval in Figure 2) but it did not increase in the old-reordered condition, as shown by the fact that the Δc_1 posterior (black bar for 95% highest- interval) was centered on approximately zero. In contrast, the old-reordered condition did produce an increase in *a*, as shown by the positive Δa_1 estimates (black bar). Against our expectation, the posterior of Δa_2 (red bar) was concentrated around zero, suggesting that the old-same condition did not increase *a*. Hence, our experimental manipulations affected the parameters even more selectively than we anticipated: Whereas we expected the old-same condition to increase both memory for elements (parameter *a*) and memory for bindings (parameter *c*), only the latter increase received support from the data.

The top panel of Figure 2 also shows the fit of the predicted to the observed proportions of responses for the unconstrained (red) and the constrained M^3 (blue). The unconstrained model fit the group means virtually perfectly, which is not surprising because it is saturated with regard to the population-level parameters (i.e., it has seven free parameters to account for seven independent mean proportions). The constrained model version that implements our expectations (i.e., $\Delta c_1 = 0$ and $\Delta a_2 = \Delta a_1$) fit the data nearly as well as the unconstrained model (ΔWAIC = 1.4), and the deviations of model predictions from the data are minimal. This shows that, although freely estimated Δa_1 and Δa_2 parameters differed, the evidence for that difference is only weak. Figure 3 presents the posteriors of the parameter values of the constrained M^3.

To conclude, Experiment 1 provides evidence for the validity of the M^3 core parameters by showing that they are selectively influenced by a manipulation that should affect only memory for elements (i.e., repeating an element in a different position) and a manipulation that affects memory for bindings (i.e., repeating an element in the same

position). We cannot conclusively say whether the latter manipulation also affects memory for elements, but we regard that question to be of secondary interest. The main goal of Experiment 1 was to test for selective influence, and that test was successful. On that basis, we consider the model sufficiently validated to warrant its further exploration.

Experiments 2-4: Complex Span

In the complex-span paradigm, encoding of memory items is interleaved with a distractor task, such as reading a set of words or working through a set of arithmetic problems. One variable that has been shown to strongly influence memory performance in complex-span tasks is the pace at which the steps of a distractor task are required (Barrouillet, Bernardin, & Camos, 2004): A slower pace enables better recall of the memory list. On the assumption that completion of each processing step takes a roughly constant amount of time regardless of pace, a slower pace implies longer periods of free time in between the processing steps. Apparently, this free time is beneficial to memory. There are currently two competing explanations for the beneficial effect of free time. One is that free time is used to increase the strength of the memory items through a process of rehearsal and/or refreshing (Barrouillet et al., 2004; Camos, Lagner, & Barrouillet, 2009), or through consolidation of the items in working memory (Bayliss, Bogdanovs, & Jarrold, 2015). The other explanation is that free time is used to remove the distractors from working memory (i.e., to unbind them from their context), thereby reducing interference from them (Oberauer, Lewandowsky, et al., 2012). Removal of distractors differs from filtering them in that filtering affects the memory strength of a stimulus during encoding, whereas removal affects it after it has been encoded (Hasher et al., 1999; Lewis-Peacock, Kessler, & Oberauer, 2018). We extended the basic model to include both these processes. The extended model versions are tailored to the design of the first three experiments, and therefore we first describe these experiments.

Design

The first three experiments used slightly different variants of a complex-span paradigm. Experiment 2 was published as Experiment 1 in Oberauer and Lewandowsky (2016); Experiments 3 and 4 have not been published before. In each experiment, participants tried to remember a list of nouns, and in between encoding of the memoranda they processed other nouns – distractors – that did not have to be remembered. The same judgment was required for memory words and distractor words: Participants had to decide whether the object that the noun referred to was larger or smaller than a soccer ball. Details of the methods of Experiments 3 and 4, as well as those results that are not the target of modelling, are reported in Appendix B.

In Experiment 2, each of five memory words (displayed in red) was followed by exactly one distractor word (displayed in black). Thus, as in a conventional complex-span task, participants knew in advance which word they would have to remember and which word was a distractor. This knowledge could be used to filter distractors, that is, to encode the distractors into working memory with reduced strength relative to the memory items, or in the extreme case of perfect filtering, not to encode them at all. The only experimental manipulation in Experiment 2 pertained to the free time following distractors: After each judgment on a distractor word, the free time was either short or long. This free time could be used to improve the strength of previously encoded memory items (i.e., through consolidation, rehearsal, refreshing, elaboration, or some other process), to remove the previously processed distractors from working memory, or for both kinds of processes.

Experiments 3 and 4 served to distinguish between filtering of distractors during encoding and removal of distractors after they have been encoded into working memory. Memory words alternated with distractor words in a random fashion, so that participants did not know in advance whether the next word would have to be remembered or not. The status

of each word – as a memory word or a distractor – was indicated by a cue for each word. In pre-cue blocks the status cues preceded each word, so that participants could still filter distractors. In post-cue blocks the status cues followed the size judgment on each word, so that distractors could not be filtered during encoding up to the point when the size judgment was completed. In both cueing conditions, distractors could be removed from working memory after finishing the size judgment. In Experiment 3, the free time following each distractor was varied, thereby giving participants a short or a long time interval for removing the preceding distractor, or to strengthen the memoranda.

In Experiment 4, the free time after each memory item was varied instead. This free time can be used for strengthening the memoranda, but it is less straightforward to use it to remove distractors. In the SOB-CS model of complex span, which uses distractor removal to reduce distractor interference (Oberauer, Lewandowsky, et al., 2012), distractors can be removed only immediately after having been encoded, because SOB-CS needs to have a representation of the to-be-removed content in the focus of attention. Thus, if free time follows encoding of a memory item, that item – rather than any previously encoded distractor – will be in the focus of attention, and as a consequence, the model could not remove distractors from working memory. It is conceivable, however, that the removal mechanism in SOB-CS is too constrained, and that distractors preceding the last-encoded memory item can also be removed in the free time after that item. We will therefore allow for both strengthening the memory items and removal of distractors during free time in the models for all three experiments.

Measurement Models

The extended measurement models include two parameters reflecting potential processes during the free time following processing of a memory item or a distractor. First, we introduce the possibility that the strength of item activation a or of item-to-context binding

c of memory items – but not distractors – is increased during these free-time periods through some process of *extended encoding*. Extended encoding refers to strengthening of memory representations after their initial encoding; initial encoding occurs during stimulus presentation, whereas extended encoding proceeds in the absence of the stimulus. We remain neutral on what exactly extended encoding entails – it could be rehearsal, refreshing, consolidation, or elaboration. We model this process as a linear increase of *a* and/or *c* over time with slope *e*. The choice to model extended encoding by a linear growth was made only for convenience given that we have only two time points to estimate that growth function; with more time points, different growth functions could be compared to each other in the future.

Second, we introduce the possibility that the strength of activation *a* or of bindings *c* of distractors – but not of memory items – is reduced during the free time through some process of gradual *removal*. We model this process as an exponential decline of *a* and/or *c* over time with rate *r*. An exponential decline towards zero (as opposed to a linear decline) is necessary to model removal – even if only two time points are available – because continuing removal should never push strength below zero. As noted earlier, the zero point on the ratio scale of memory strengths is interpreted as the absence of information in memory.

The extended model equations are:

$$\begin{aligned} A(correct) &= b + (1 + et_c)(a + c), \\ A(other\ item) &= b + (1 + et_c)a \\ A(distractor\ \ in\ position) &= b + \exp(-rt_c)f(a + c) \\ A(other\ distractor) &= b + \exp(-rt_c)fa \\ A(NPL) &= b. \end{aligned} \tag{5}$$

Here, t_c is the free time following each distractor (Experiments 2 and 3) or each memory item (Experiment 4) in condition *C*; *e* is the rate at which memory strength increases

over time through extended encoding; *r* is the rate of exponential removal of distractors over time; and *f* is the filtering parameter for distractors as explained earlier. In Experiments 3 and 4, the filtering parameter is used only for the pre-cue condition; in the post-cue condition, *f* is set to 1 because people cannot filter during encoding as they do not know whether a stimulus is a memory item or a distractor.

The equations above represent the model version in which extended encoding and removal affect both activation and context-binding strength. We also consider model versions in which extended encoding applies to activation *a* only, or to binding strength *c* only, or to neither of them, and model versions in which filtering, or removal, applies to persistent activation only, to binding strength only, or to neither. Fully crossing all these model variations results in 4 x 4 x 4 = 64 model versions, all of which were applied competitively to the data of Experiments 2 to 4.

Bayesian Hierarchical Implementation

The Bayesian hierarchical measurement model for complex span is specified by the following equations. In the equations below "=" denotes a fixed value assignment, whereas "~" denotes that the variable on the left-hand side is distributed according to the distribution on the right-hand side.

$$
\begin{aligned}
&\mathbf{y}_{i,c,\cdot} \sim Multinomial(\mathbf{P}_{i,c,\cdot}, N_{i,c}) \\
&P_{i,c,j} = \frac{n_j A_{i,c,j}}{\sum_{k=1}^{K} n_k A_{i,c,k}} \\
&A_{i,c,1} = 0.1 + (1 + e_i t_c)(c_i + a_i) \\
&A_{i,c,2} = 0.1 + (1 + e_i t_c) a_i \\
&A_{i,c,3} = 0.1 + \exp(-r_i t_c)\, f_i (c_i + a_i) \\
&A_{i,c,4} = 0.1 + \exp(-r_i t_c)\, f_i a_i \\
&A_{i,c,5} = 0.1
\end{aligned}
\tag{6}
$$

The first set of equations above describes the bottom of the hierarchy. The data **y** of each individual *i* in condition *c* are represented as a vector of response frequencies over the five response categories (correct, other item, distractor in position, other distractor, and NPL). This frequency vector is described by a multinomial distribution with a probability vector **P** over the five categories, and the total number of observations *N*. The probabilities of each response category, $P_{i,c,j}$, are obtained by normalizing the activation values across the K=5 categories using Luce's choice rule (Eq. 2). In this normalization step each response category *j* needs to be weighted by the number of candidates in that category, n_j, because participants' choices are choices among individual candidates, not categories. Specifically, while there is only 1 correct item, there are, for example, 5 NPL's, and Luce's choice rule must take this imbalance into account.

Activation values for each response category are computed according to the model equations given above, using the parameters for each individual *i,* as described in the set of Equations (7) below. On the next level of the hierarchy, individual-level parameters are drawn from distributions specified by population-level parameters (Equations 8). At this point we need to determine the scale on which we measure individual differences in parameter values. The model parameters are not defined on the full real line (i.e., none of them can reasonably be negative, and the filter parameters are constrained between 0 and 1), and therefore they cannot be normally distributed. At the same time, it is convenient to describe individual differences by a normal distribution of parameter values over individuals, which implies that parameter values are measured on a real-valued scale. Doing so enables describing effect sizes in terms of standard deviation units, as in Cohen's *d* statistic for effect sizes, and facilitates estimating the correlations between parameters. One way to resolve this tension is to use normally distributed variables to describe individual differences, and transform them into the actual parameter values through a non-linear function. We used this solution for the

filter parameter, *f*, which is generated from a normally distributed variable ζ through a logistic transformation. For the remaining parameters we drew the individual parameters from a normal distribution without transformation. The normal distribution is not meant to describe the true distribution of parameters but to function as a convenient approximation of the true distribution.[5]

$$\begin{aligned} c_i &\sim Normal(\mu_c, \sigma_c) \\ a_i &\sim Normal(\mu_a, \sigma_a) \\ e_i &\sim Normal(\mu_e, \sigma_e) \\ r_i &\sim Normal(\mu_r, \sigma_r) \\ \zeta_i &\sim Normal(\mu_f, \sigma_f) \\ f_i &= \frac{1}{1+\exp(-\zeta_i)} \end{aligned} \tag{7}$$

Finally, we set moderately informative normal priors on all population-level means, and Gamma priors on the standard deviations:

$$\begin{aligned} \mu_c &\sim Normal(20,10) & \sigma_c &\sim Gamma(1,0.01) \\ \mu_a &\sim Normal(2,10) & \sigma_a &\sim Gamma(1,0.01) \\ \mu_e &\sim Normal(1,10) & \sigma_e &\sim Gamma(1,0.01) \\ \mu_r &\sim Normal(1,10) & \sigma_r &\sim Gamma(1,0.01) \\ \mu_f &\sim Normal(0,10) & \sigma_f &\sim Gamma(1,0.01) \end{aligned} \tag{8}$$

To summarize, the model estimates the memory strength parameters for each individual participant, but the estimates were constrained to be samples from a parent distribution whose mean and variance was also estimated.

Results Experiment 2. Figure 4 illustrates the distribution of the ΔWAIC values from Experiment 2 over the 64 models. The darker a square, the worse is a model's fit in

[5] Sampling of negative values from the Normals could be avoided by truncating them at zero, but in practice we never encountered the need to do this with the present models.

comparison to the best-fitting model. The best-fitting model, by definition, has a white square with a WAIC difference of zero from itself. The bar graphs in Figure 5 reflect the behaviorally observed probabilities of choosing a response in each of the five response categories. The empirical probabilities are calculated as the number of responses in each category divided by the number of response candidates in each category. For instance, although the absolute number of *distractor in position* responses was smaller than the number of *other distractor* responses, the probability of choosing each individual other distractor was smaller than the probability of choosing each *distractor in position*, because there were more *other distractors* in the test array than there were *distractors in position.*

The model with the best fit had an extended-encoding parameter affecting only the binding strength *c*, a removal rate parameter also affecting only *c*, and a filter parameter affecting both *a* and *c*. The red dots in Figure 5 represent the means of the posterior predictives of this model. The posterior predictives are samples of predicted data obtained from the M^3 with parameter values sampled from their posteriors. Figure 6 shows the posteriors of the parameter means across participants. These distributions are obtained by averaging the posterior parameter values of all individuals at each MCMC sample, thereby generating a sample of their posterior mean. The posterior means of the best-fitting model versions across all experiments are also summarized in Table 1.

Several conclusions can be drawn from these results. First, longer free time after distractors increased the number of correct responses at the expense of all error categories (perhaps with the exception of NPLs). The model attributes this effect to extended encoding of item-context bindings, and to removal of distractor-context bindings. The effect of extended encoding on *c* was substantial: We can calculate the increase in binding strength *c* through extended encoding for each free-time condition *C* as $\Delta c = c \cdot e \cdot t_c$. Using the means of the posteriors for each parameter, we obtain $\Delta c = 13.4 \times 0.59 \times [0.2, 1.5] = [1.6, 11.9]$. In

other words, the long free time of 1.5 s nearly doubled the estimated strength of item-context bindings.

Second, distractors are encoded into working memory, but with only about half the strength compared to memory items, as reflected in the filtering parameter. In addition, distractor-context bindings – but not distractor activations – are removed after encoding. The high removal rate implies that removal of bindings proceeds very rapidly: With $r = 16.7$, the strength of distractor-context bindings is reduced by 96% after 0.2 s of free time, and virtually eliminated after 1.5 s. In this way, the model predicts the pattern of distractor selections: Because distractor activation is not reduced through removal, distractors are chosen as responses much more often than NPLs at both free-time levels; because distractor-context bindings are rapidly and strongly removed, the probability of distractor intrusions are predicted to be only slightly higher for *distractors in position* than for *other distractors*.

The rapid removal rate implies only a negligible difference in distractor-context bindings between the short and the long free-time conditions of this experiment. Therefore, rapid removal is difficult to distinguish from particularly strong filtering of distractor-context bindings at encoding. A further model version, in which separate filtering parameters f_a and f_c were applied to a and c, respectively, provided a slightly better fit to the data of Experiment 2, with an estimate of mean $f_c = 0.03$. To resolve this ambiguity, in Experiments 3 and 4 we included a condition in which distractors were identified as such only after they have been encoded into working memory. In this post-cued condition, any reduced strength of distractors relative to memory items must be attributed to removal after initial encoding. To foreshadow, Experiments 3 and 4 will confirm that distractor-context bindings are rapidly removed after encoding.

Experiments 3 and 4. For both experiments, the model with a filtering parameter applied to both a and c, extended encoding applied to a and c, and removal applied to c only

fit the data best. Figures 7 and 10 show the WAIC differences for the models applied to Experiment 3 and 4, respectively. The probabilities of responses together with the model predictions are presented in Figures 8 and 11, for Experiments 3 and 4, respectively. The posteriors of parameter means are plotted in Figures 9 and 12. The results of both experiments are remarkably consistent with each other. The parameter estimates for *c* and *a* roughly match those for Experiment 2; the effect of extended encoding was estimated to be even stronger than in Experiment 2 and affected not only item-context bindings but also item activation. The filtering parameter estimates approximately match the *f* parameter of Experiment 2.

The removal parameter again implies a very rapid removal process operating only on *c*. Strong and rapid removal is necessary for the model to explain the relative proportions of selections of memory items and of distractors in the post-cued condition, in which the strength of distractors could only be reduced by removal. Although post-cued distractors intruded into recall more often than pre-cued distractors, the difference was modest, and post-cued distractors were still chosen much less frequently than memory items, implying that their strength must have been reduced substantially in response to the post-cue. Because the post-cues were presented only after the distractors have been presented and processed (i.e., after participants made their size judgment on them), this reduction in strength can only be attributed to removal. The large estimates of the removal rate *r* imply that the removal of distractor-context bindings is nearly complete even after a short free-time interval: The proportional reduction of *c* through removal, estimated from the means of the parameter posteriors, was about 98% for short, and essentially 100% for long free time.

Discussion

Besides demonstrating the usefulness of the M^3 framework, several substantive conclusions can be drawn from Experiments 2 to 4. First, replicating previous experiments (Oberauer, Farrell, et al., 2012), distractors intruded more often than NPLs. This demonstrates

that distractors are encoded into working memory to some extent, confirming an assumption in the SOB-CS model (Oberauer, Lewandowsky, et al., 2012). The measurement models provide a more detailed picture of this process. Estimates of the filtering parameter show that words known to be distractors during encoding are encoded with about half the strength of the memory items. This filtering applies equally to activation and binding.

After encoding, distractors are to some extent removed from working memory. Whereas the existence of such a removal process is predicted by the SOB-CS model, the details that emerge from the present experiments do not agree well with that model. In SOB-CS, removal consists of the gradual untying of distractor-context bindings. The measurement models instead reveal a very rapid removal of distractor-context bindings. The M^3 predicts this removal to be virtually complete after 1.5 s of free time. This might not be entirely accurate: In all three experiments, there was still a somewhat higher probability of recalling a distractor close to the current list position than another distractor even in the long free-time condition; a tendency that the model predictions miss. We explored whether the model prediction could be improved in this regard if removal of distractor-context bindings were modelled as a rapid exponential decline to an above-zero asymptote. We extended the best-fitting model version for Experiments 3 and 4, respectively, by adding the lower asymptote of removal as a further free parameter. This extension improved the model fit slightly for Experiment 4 (ΔWAIC = 15), but not for Experiment 3 (ΔWAIC < 1).

The measurement models also afford separating the effects of extended item encoding and of distractor removal during free-time intervals: A free-time effect through removal implies a substantial reduction of distractor choices relative to NPLs, with all other response categories increasing. Extended encoding predicts an increase in correct responses (when applied to bindings) or of correct and other-item responses (when applied to memory for elements), and all other response categories declining by the same proportion. The latter

prediction matches the data better, and therefore the M^3 attributed the free-time effect primarily to an increase in memory strength through extended encoding. This was true for free time immediately following each item (Experiment 4), but also for free time following distractors (Experiments 2 and 3), implying that free time can be used to boost memory strength of previously encoded items also after a disruption by an intervening distractor. Extended encoding of memory items contributed much more to the beneficial effect of free time than removal of distractors. This finding demands a revision of the SOB-CS model, in which distractor removal alone accounts for the free-time benefit. At least for words as memoranda, memory strength does not remain constant after their initial presentation (1.7 s in the present experiments) but continues to grow during subsequent free-time periods, when the word was no longer visible. Across the three experiments, the best-fitting models assumed extended encoding to strengthen either item-context bindings alone (Experiment 2) or both bindings and item activation (Experiments 3 and 4).

Several processes have been proposed in the memory literature that could be responsible for the extended encoding benefit. First, we could assume that articulatory rehearsal boosts memory strength for rehearsed items. Tan and Ward (2008) found that serial recall of words – uninterrupted by distractors – is better when words are presented at a slower rate (5 s vs. 1 s per word). Through an overt-rehearsal procedure Tan and Ward monitored participants' articulatory rehearsal and found that they engaged in more cumulative rehearsal at the slower presentation rates. Moreover, the extent of cumulative rehearsal correlated positively with serial recall performance. These findings are compatible with the assumption that articulatory rehearsal does not only protect memory representations from decay, but rather strengthens them beyond their state after presentation (see also Nishiyama & Ukita, 2013). Against this possibility, an exploration of rehearsal mechanisms in the context of computational models of serial recall revealed some principled limitations of rehearsal

(Lewandowsky & Oberauer, 2015), and a series of experiments found that when cumulative rehearsal was experimentally increased through instruction, it had no beneficial effect on memory (Souza & Oberauer, 2018). Another possible process that could boost memory strength is refreshing, defined as attending to a representation in working memory after its presentation (Johnson, 1992; Raye, Johnson, Mitchell, Greene, & Johnson, 2007). Like rehearsal, refreshing has been invoked as a mechanism for maintaining memory strength in decay theories (Barrouillet et al., 2004), but it could also be conceptualized as a process that increases memory strength above the level reached after initial encoding. Evidence for that possibility comes from a study of visual working memory (Souza, Rerko, & Oberauer, 2015): Asking participants to attend to a subset of items in a memory array improved memory for those items. This effect of refreshing is also found for verbal materials, though weaker than for visual and spatial materials (Souza, Vergauwe, & Oberauer, 2018). However, at present it is unclear whether people spontaneously refresh memory representations during free time (Vergauwe et al., 2016; Vergauwe, Langerock, & Cowan, 2018).

A third possible interpretation of the extended encoding process is as consolidation. Consolidation can be distinguished from initial encoding in that consolidation continues to operate after a mask has erased sensory information from the stimulus (Nieuwenstein & Wyble, 2014; Ricker & Cowan, 2014). Whereas earlier investigations of "short-term consolidation" of information in working memory estimated it to be complete after less than a second (Jolicoeur & Dell'Acqua, 1998), subsequent research suggests that consolidation can continue for a longer time (Bayliss et al., 2015; Nieuwenstein & Wyble, 2014). However, consolidation is assumed to be interrupted by a processing demand that requires central attention, or by encoding of another item (Ricker & Cowan, 2014), and it is not clear whether consolidation can resume after such an interruption.

A fourth interpretation of extended encoding is as elaborative rehearsal (Craik & Watkins, 1973), defined as creating a richer semantic representation of the memory material by relating items to each other or to knowledge in long-term memory. Elaboration is known to improve episodic long-term memory (Craik & Tulving, 1975), but so far there is no evidence that it also improves recall from working memory (Bartsch, Singmann, & Oberauer, 2018). To conclude, the present experiments provide compelling evidence that some process continues to strengthen item-context bindings, and perhaps also item activation, during free time long after their initial encoding. The nature of this process is not clear, and certainly deserves further investigation.

Experiment 5: Working Memory Updating

In the memory-updating paradigm, an initial set of memory items is updated by replacing individual items with new items. The paradigm can be traced back to the early days of experimental cognitive psychology (Yntema & Mueser, 1962). In individual-differences studies, updating tasks are among the best measures of working-memory capacity (Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000; Wilhelm, Hildebrandt, & Oberauer, 2013). Here we used a version closely modeled after Kessler and Meiran (2006). Participants encoded an initial set of four words displayed across a row of four frames, and subsequently replaced them with new words that are displayed one by one in individual frames in a random sequence. At the end participants try to recall the last word that was presented in each frame.

In a previous series of experiments using this task in a self-paced mode, Ecker and colleagues were able to separate two processes involved in working-memory updating: Removal of the old representation and encoding of the new one (Ecker, Lewandowsky, & Oberauer, 2014; Ecker, Oberauer, & Lewandowsky, 2014). In these experiments, each new stimulus was preceded by a cue indicating the frame in which the stimulus will appear. This cue informed participants about which old item in the current memory set will be replaced

next, but it did not reveal or predict the new item. Participants appear to use the time between the cue and the new memory item to remove the old item. This is shown by two findings: First, after a longer cue-stimulus-interval, participants took much less time to update memory once the new stimulus was given. Second, if the new stimulus was similar – or even identical – to the old stimulus it replaced, updating times were faster, but this similarity advantage was eliminated by a long cue-stimulus interval (Ecker, Lewandowsky, et al., 2014; Ecker, Oberauer, et al., 2014). Our aim for Experiment 5 was to apply the M^3 framework to the updating paradigm and to extend it by parameters measuring the removal of old items and the encoding of new items.

Design

The M^3 framework is so far applicable only to response choice data, not to response times. Therefore, we did not use the self-paced updating task of Ecker and colleagues, but a computer-paced version of the same task. After presentation of the initial four memory words – displayed from left to right across a row of four frames – participants saw a series of further words displayed one by one in randomly selected frames. Each new word was presented for 0.5 s. For each new word participants had to replace the word they remembered for that frame by the new word. Each new word was preceded by a cue indicating in which frame it would appear (a fixation cross in the center of the frame, displayed for 0.2 s). The length of the series of new words was unpredictable so that participants had to expect to be tested at any moment – this served to discourage them from not attending to the earlier words in the series. The design crossed a variation of the cue-word interval (0.1 vs. 1.1 s after the offset of the cue) with a variation of the interval between presentation of a new word and the onset of the next cue in another frame (word-cue interval, 0.1 vs. 1.1 s after offset of the word). The purpose of these manipulations was to vary the time for removal of the old item (cue-new word interval) and the time for encoding of the new item (new word-next cue interval).

At the end of a series of updating steps, memory was tested by presenting participants with an array of 12 words – the four last words in each frame, the four next-to-last words (which we will refer to as "old words"), and four NPLs. The four frames were probed in a random order, and participants were asked to select the last word for the probed frame from the 12 candidates. We sorted their responses into five categories: Correct words, other words from the set of last words, old words in the probed frame, other old words, and NPLs. Procedural details of the experiment are provided in Appendix B.

Measurement Models for Memory Updating

We extended the basic model by three processes. First, as for the complex-span models, we included an extended-encoding process to capture the increase of memory strength for both activation and item-context bindings during free time t_e following presentation of a new word. Second, we included a process of removing the old item after it has been identified by the cue. Removal was modelled as an exponential reduction of memory strength during the removal interval t_r. Third, we include a parameter d that reflected the proportional weakening of the old item through encoding of the new item. This latter process differs from removal in that it is not time dependent: The same proportional reduction of c and a of old items is assumed for all experimental conditions. One motivation for introducing this instantaneous deletion process was the consideration that an old item might be much more easily removed from working memory when being replaced by a new item than during the cue-word interval, when no new item is yet available. Some evidence for that possibility comes from the directed-forgetting literature (Hertel & Calcaterra, 2005). Another motivation for the d parameter was simply that without it, none of the measurement models fit the data well. The model equations for the memory-updating experiment are:

$$
\begin{aligned}
&A(correct) = b + (1 + et_e)(a + c),\\
&A(other\ item) = b + (1 + et_e)a,\\
&A(old\ in\ position) = b + \exp(-rt_c)d(1 + et_e)(a + c),\\
&A(other\ old) = b + \exp(-rt_c)d(1 + et_e)a,\\
&A(NPL) = b.
\end{aligned}
\tag{9}
$$

These equations are in most regards analogous to those for complex span, with one important difference: Old items – which take the role of distractors in the updating paradigm – are originally encoded in the same way as new items. Hence, in a condition with longer time for encoding new items, longer time was also available for encoding the old items on a previous updating step. Therefore, the extended-encoding term applies not only to the current but also the old items.

As before, the three processes – extended encoding, gradual removal, and instantaneous deletion – can each affect only *a*, only *c*, both *a* and *c*, or neither of them. Crossing these four possibilities for each of the three processes yields $4^3 = 64$ model versions. A further dimension of model variation comes from the following consideration: When designing the experiment, we hoped that participants would use the cue-word interval only for removing the old item, and use the word-cue interval only for encoding the last-presented word. However, participants might choose to use these intervals otherwise: First, they could use the word-cue interval to continue removing the old item in the current frame while encoding the new item. Second, they could use the cue-word interval to continue with extended encoding of the new word from the preceding updating step (in another frame), instead of, or in addition to, removing the old word in the cued frame. To accommodate these possibilities, we created four model variants differing in how the time for removal t_r, and the time for extended encoding t_e, were defined (using CWI to refer to the cue-word interval, and WCI for the word-cue interval):

(1) $t_r = 0.2 + CWI$; $t_e = WCI$ (10)

(2) $t_r = 0.2 + CWI + 0.5 + WCI$; $t_e = WCI$

(3) $t_r = 0.2 + CWI$; $t_e = WCI + 0.2 + CWI$

(4) $t_r = 0.2 + CWI + 0.5 + WCI$; $t_e = WCI + 0.2 + CWI$.

Here, we added the 0.2 s of cue presentation to the time for removal in all variants, and the 0.5 s of word presentation to the time for removal in those versions assuming that removal continues after presentation of the new word. We also added the 0.2 s of cue presentation to the time for extended encoding because when extended encoding of the previous word continues into the CWI of the next updating step, it also continues during presentation of the cue. We did not add the word presentation time to t_e because t_e reflects only the time for extended encoding after stimulus offset.

Crossing this variation of time definitions with the other model variations generates a total of 256 model versions. All models were implemented as Bayesian hierarchical models in the same way as those for complex span.

Results

The model with the best fit according to WAIC used time definition (4), implying that both cue-word interval and word-cue interval were used for removal of old items and extended encoding of new items. Extended encoding affected both a and c, whereas removal and deletion affected only c. Figure 13 plots the WAIC differences. Figure 14 shows the probabilities of choosing a response in each of the five categories, together with the model predictions. Figure 15 shows the posterior means of parameter estimates across participants.

A number of conclusions can be drawn immediately from these results. Performance increased with longer cue-word intervals and with longer word-cue intervals; the effects of the

two intervals did not differ appreciably. This is why a model treating the two intervals interchangeably fit the data best. Old items were selected much more often for recall than NPLs, implying that old items were not completely removed from working memory. Old items in the tested frame were much more likely to be recalled than other old items – and even more likely than other current items. This shows that bindings of old items to their contexts remained strong after updating.

The parameter estimates for c and a are much lower than those for the complex-span experiments, implying that both item activation and item-context bindings were weaker in this experiment. This could reflect a feature of the updating paradigm, or the fact that in this experiment we gave participants a much shorter presentation time for encoding each word (0.5 s compared to 1.7 s in the complex-span experiments). The small values of c and a after initial encoding are substantially increased by extended encoding. The time for extended encoding was 0.4 s in the condition with two short intervals, 1.4 s in the conditions with one short and one long interval, and 2.4 s in the condition with two long intervals. With mean e = 4.3, these result in c = 5.4, 14.0, and 22.6, respectively, and in a = 0.30, 0.77, and 1.25. These values are comparable to those found with complex span.

The removal parameter implies a fairly slow gradual weakening of item-context bindings for old items. The removal times for the four experimental conditions were 0.9 s (both short intervals), 1.9 s (one short, one long interval), and 2.9 s (both long intervals). With a mean removal rate of 0.53, the bindings of old words would be reduced by factors of 0.62, 0.37, and 0.22 for the three removal times, respectively. In addition, rapid deletion – as reflected in the d parameter – reduced binding strength of old items further by about one half. This still leaves substantial item-context bindings for old items, which explains the high probability of responding with an old item in the tested frame. In contrast to bindings, the activation of old items is not reduced at all relative to those of new items. As a consequence,

old items in other than the tested frame are chosen about as often as current items in other frames. As both these error types are rare, not much damage is done by leaving the fairly low strength of memory for individual items – reflected in the small *a* parameter – unchanged through updating.

Discussion

Our motivation for independently varying the cue-word interval and the word-cue interval was to separately influence the time for removal of old items and the time for encoding of new items. This did not work out, because participants used both time intervals for both processes. In this regard our computer-paced version of the updating task is probably different from the self-paced version of Ecker and colleagues (Ecker, Lewandowsky, et al., 2014; Ecker, Oberauer, et al., 2014). In the self-paced version, participants are instructed to move on to the next updating step only when they have completed the preceding step, so that it is unlikely that they use the cue-word interval to continue extended encoding of the new word from the preceding updating step. In the computer-paced version, by contrast, the next updating step often commences before participants feel that they have sufficiently encoded the word from the previous step, and therefore use some of the following cue-word interval for extended encoding. Yet, the cue-word interval is also used for removing the item-context bindings of the old item in the cued frame. We cannot say whether extended encoding and removal proceed in parallel, or whether people switch back and forth between both processes during both intervals. Switching between two serial processes would imply that the *e* and *r* parameters underestimate the true rate of extended encoding and removal, respectively, because each of these processes would actually only use half the available time, on average.

Removal of bindings of old items was found to rely on two processes – one gradual proceeding with removal rate *r*, and one very rapid, represented by the proportional deletion parameter *d*. The fast removal process is reminiscent of the findings with complex span

(Experiments 2 to 4), although it is substantially less complete, leaving room for the second, much slower removal process to reduce old item-context bindings further. Despite being weakened considerably through removal, bindings of old items to their frames remain in working memory after updating. They generate proactive interference in the form of intrusions of old items in the tested frame. The conclusion that old item-context bindings linger on after an updating step converges with previous research (Oberauer & Vockenberg, 2009). That said, in the present updating paradigm, intrusions of old items could also reflect occasional trials on which participants failed to update at all because they did not have sufficient time. When they missed the last updating step in a frame completely, they would select the next-to-last item (i.e., the old item) for that frame when tested. It is unlikely that complete updating failures explain all of the intrusions of old items in the tested frame, because these intrusions occurred at a rate much higher than other errors even in the condition with long cue-word and word-cue intervals, in which there were 2.9 s for each updating step.

Experiment 6: Age Differences in Working Memory

One goal of developing measurement models is to use them as tools for measuring individual differences or group differences in latent variables, such as the strength of item-context bindings, or the rate of extended encoding. Experiment 6 serves to illustrate the use of the M^3 framework for investigating age differences in working memory. We administered the complex-span task of Experiment 2, and the memory-updating task of Experiment 5, to samples of healthy young and old adults (see Appendix B for details on participants and methods). It is well established that old adults perform more poorly than young adults on tests of working memory (Bopp & Verhaeghen, 2005; Salthouse, 1994). Through the lens of measurement models we can ask which latent variables, as represented by the parameters of the model, differ in what way between young and old adults.

Results

We applied the same set of measurement models as before to the data from each task. Each model was applied simultaneously to data from both age groups, estimating different population-level parameters for young and old adults. The model fit was evaluated for both age groups jointly.

Complex Span. As for Experiment 2, the best-fitting model for complex span was the one with a filtering parameter affecting both parameters *a* and *c*, and extended encoding and removal affecting only *c*; WAIC differences are shown in Figure 16. The probabilities of choosing a response from each category are shown in Figure 17 together with the predictions from that model; Figure 18 shows the posteriors of the parameter means for each age group.

The data of the young group replicate Experiment 2 in all respects. As expected, older adults performed worse, and in particular committed more errors choosing another list item, or another distractor. By contrast, their tendency to choose a *distractor in position*, or an NPL, was not markedly elevated. The measurement model helps to explain this age effect on error patterns: Older adults were estimated to have much reduced strength of item-position and distractor-position bindings (parameter *c*), whereas their ability to remember individual items and distractors regardless of position (parameter *a*) was unimpaired. As a consequence, old and young adults were equally able to distinguish memory items from NPLs, but older adults were less able to place the memory items in their correct positions, thereby often confusing the correct item with other list items.

The age differences in parameter estimates can best be assessed through the posteriors of the age difference in the population-level means, plotted in Figure 19. These plots also indicate what percentage of the posterior probability density falls on each side of zero. These values can be interpreted as the posterior probability that older adults have a smaller (or

larger) parameter value than young adults. For instance, older adults have a smaller *c* parameter with a probability approaching 1 (after rounding), whereas their *a* parameter is larger than that of young adults with a probability of .98. There was no evidence for an age difference in filtering out or removing distractors (parameters *f* and *r*, respectively), and no evidence that older adults are less efficient in further strengthening bindings through extended encoding (parameter *e*).

Memory Updating. The WAIC difference values from the model comparison are shown in Figure 20. As for Experiment 5, the best-fitting model used time definition (4), by which both cue-word interval and word-cue interval were used for removal of old items and extended encoding of new items. Extended encoding and deletion affect only *c*, and removal only *a*. In these regards the best-fitting model differs from that for Experiment 5, in which extended encoding affected *a* and *c*, and removal affected only *c*. Both *c* and *a* of young adults were much larger than in Experiment 5, probably due to the substantially longer presentation time of words, which afforded longer encoding – the values obtained here are similar to those found for complex span, which had a comparable encoding duration (i.e., the average size-judgment times were about 1.1 s, see Table A1). Figure 21 shows the empirical probabilities of choosing a response from each category, together with the model predictions; Figure 22 presents the posterior means of parameter estimates for the two age groups.

Different from Experiment 5, the rapid removal process captured by parameter *d* was nearly complete for the young adults in the present experiment. This left very little to do for the slower removal process reflected in parameter *r*; this parameter was estimated close to zero, rendering it virtually ineffective. The more thorough rapid removal of old items in the present experiment compared to Experiment 5 explains why here the probability of erroneously selecting an old item was much reduced compared to the previous experiment, whereas the probability of selecting a current item from another than the tested position was

not. This comparison between the two updating experiments lends support to our assumption that the rapid removal indexed by *d* is a side effect of the initial encoding the new word – with a longer presentation time initial encoding is more thorough, and with it, removal of the to-be-replaced old word is more thorough too.

Comparison of the two age groups revealed a pattern very similar to that for complex span (see Figure 23 for posteriors of age differences in population-level parameters): Compared to young adults, old adults formed weaker item-context bindings, as reflected in their smaller *c* parameters. In contrast, they were at least as effective in generating and maintaining representations of individual items, as reflected in their larger *a* parameters. Old adults were also less effective in rapidly removing old information from working memory, as shown by their larger *d* parameters, and they were less efficient in boosting item-context bindings through extended encoding (parameter *e*).

Discussion

Whereas the complex-span experiment replicated all findings from Experiment 2, the memory-updating experiment yielded a different pattern of errors. As a consequence, the best-fitting model differed in some regards from that for Experiment 5, and the parameter estimates differed as well. These differences were most likely due to the longer word presentation times in the present updating experiment. The difference is unlikely to be merely a difference in the total available time for an updating step: The condition with short CWI and long WCI in Experiment 5 had a longer total time for each updating step than the condition with short CWI and short WCI in the present experiment (1.9 vs. 1.6 s), and yet the probability of choosing an old item – in particular the old item in the updated position – was larger in the *short-long* condition of Experiment 5 than in the *short-short* condition of young adults in Experiment 6 ($p = .16$ vs. 0.09). The duration for which the new word is actually present on the screen appears to be more important for rapid removal of the old item than the

total duration of the updating step. Future research with the updating paradigm could vary the presentation duration within an experiment and investigate how it affects the model parameters.

The primary interest of Experiment 6 was on age differences in the latent variables measured by the model parameters. The results are consistent across both experimental paradigms: Old adults were at least as good as – probably better than – young adults in generating and maintaining strong representations of individual elements, but they were impaired in creating and maintaining content-context bindings. Although old adults appeared to be less efficient in strengthening bindings through extended encoding in the updating task, no such age effect was observed in the complex-span task. Moreover, there was no age difference in filtering distractors. There was also no age difference in the removal of distractors in complex span, but old adults were somewhat less effective in removing old item-context bindings in the updating task.

Model Recovery and Parameter Recovery

Measurement models in the M^3 framework can be used for two purposes: Finding the model variant that best explains the data, and measuring parameter values for that model. We next investigate through simulation how well the measurement models for complex span and for memory updating are suited for these purposes. For each class of model we ran two sets of simulations. The *model recovery* simulations generated data from a large set of model variants, and competitively fit each model variant to each data set in the same way as we fit the model variants to the experimental data above. Good model recovery means that the correct model variant – the one that generated the data – wins the competition most of the time. The *parameter recovery* simulations generated data from the best-fitting model version, but varying the parameter values over a large, plausible range. The same model version was fit to each data set generated under the different parameter values. Good parameter recovery

means that the parameter estimates match the true parameter values from which the data were generated.

Complex Span Models

Model Recovery. Model variants were generated by varying the application of filtering, extended encoding, and removal, each ranging over four levels (none, affecting *a* only, affecting *c* only, affecting both *a* and *c*). The full combination of these variables generates 64 model variants. To save computation time, we generated data only from the 27 model variants in which each process affected *a*, affected *c*, or affected both. For each model variant we simulated data from N=30 subjects for the design of Experiment 2, with 100 trials per free-time condition. The parameter means on the group level were set to those estimated for the best-fitting model version in Experiment 2, and the standard deviations were set to ¼ of the group mean, with the exception of the filtering parameter *f*, for which we first computed the mean logit(*f*), generated normally distributed data for each subject with SD=1, and back-transformed the logit(*f*) values to individual *f* values. Each simulated data set was fit with all 64 model versions, which we compared through WAIC. The entire model-recovery procedure as described above was repeated for 20 runs.[6]

Table 2 presents the results: For each data-generating model version (i.e., the model representing the ground truth for the simulation) the table reports the proportion of simulation runs (out of 20) in which that model version won the competition (i.e., the hit rate). We also report the proportion of cases in which each model version won the competition when it was *not* the true model (i.e., the false alarm rate). The false-alarm rate helps to identify model versions that tend to win the competition unduly often because of excess flexibility not compensated for by the WAIC. Further, the table presents the average differences in WAIC

[6] To make this computationally feasible we reduced the number of MCMC steps to 10,000 (after 5000 burn-in steps).

between the true model and the best-fitting model in those cases where the true model did not win the competition. These values reflect by how much, on average, the true model lost the competition when it did. They provide an assessment of the distribution of likely values of WAIC differences between the best and the next-best model (ΔWAIC) when the true model fails to be recovered as the winning model. Any ΔWAIC substantially exceeding these values is unlikely to reflect a case in which the wrong model won the competition due to noise in the data.

Table 2 shows that not all model versions are recovered well – for some of them the hit rate was as low as .1. More comforting is the result that the best-fitting model version of Experiments 2 and 6 had a recovery hit rate of .7, with a small false-alarm rate of .05. Moreover, when a model was not recovered well, it missed winning the competition by typically not more than 3 WAIC points. For example, the first model in the table was recovered correctly only 15% of the time; however, on average it missed being picked by a WAIC difference of only 1.86. It follows that when a model version wins over other model versions by a WAIC difference much larger than 3, we can be confident that the selected model did not win the competition merely due to noise. For comparison, the ΔWAIC for the complex-span models applied to Experiments 2, 3, 4, and 6 were 7.5, 15.2, 9.7, and 9.7, respectively.

Parameter Recovery. We generated data from the best-fitting model version for Experiments 2 and 6: The version in which filtering applied to both *a* and *c*, whereas extended encoding and removal applied only to *c*. We ran five parameter-recovery experiments; each experiment varied one parameter over 8 values, keeping the other four parameters fixed at their best-fitting values from Experiment 2. The parameter values were varied as follows:

μ(c) = [5, 7, 10, **12**, 16, 20, 30, 40],

$\mu(a)$ = [0.3, 0.5, 0.7, **1.0**, 1.5, 2.0, 2.5],

$\mu(f)$ = [0.1, 0.2, 0.35, 0.5, **0.65**, 0.8, 0.9, 1.0],

$\mu(r)$ = [2, 3, 5, 8, 10, 14, **18**, 24],

$\mu(e)$ = [0.2, 0.3, 0.4, 0.5, **0.6**, 0.8, 1.0, 1.5].

These values were chosen to span a large range of reasonable values around the best-fitting values for each parameter. (The values closest to the posterior group means in Experiment 2 are printed in bold.) The SD for individual differences was set to ¼ of the group mean, except for *f*, which was generated by logistic transformation of a normally distributed logit(*f*) with SD = 1.

For each set of parameter values we generated data from N=30 subjects for the design of Experiment 2, with 100 trials per free-time condition. The data were fit by the same model version used for generating them. Figures 24 to 28 present the results, one for each of the five parameter-recovery experiments. Each figure contains five panels, showing the effect of the manipulated parameter on each of the five estimated parameters. The true value of the manipulated parameter is given on the x-axis; the estimated parameter values and their 95% HDIs are plotted on the y-axis. Optimal recovery – visualized by the dotted lines – would mean that the manipulated parameter varies with the manipulation, whereas the other four parameters are unaffected by it, implying that there is no trade-off between parameters. The figures show that this was the case, confirming that the model recovered its parameters reasonably well. In addition, in most cases the 95% HDI included the true parameter value, as should be expected if the posterior densities adequately reflect the degree of uncertainty about the true parameter values.

Memory Updating Models

Model Recovery. Model variants were generated by varying the application of extended encoding, removal, and rapid deletion over three levels each (applied to *a* only, applied to *c* only, and applied to both *a* and *c*). This was crossed with two levels of time allocation, one for the experimenter-intended allocation (time for removal = CWI and time for encoding = WCI), the other for the empirically best supported one (both CWI and WCI are used for both extended encoding of the previous stimulus and removal of the current old stimulus). Data were generated for the design of Experiment 5 with N=30 and 100 trials per timing condition using the group-mean parameter values as estimated in Experiment 5, and SD set to ¼ of the mean. The entire suite of 256 model versions (varying the application of extended encoding, removal, and deletion over four levels – including "none" – and varying time allocation over four levels) was fit to each generated data set.[7] Table 3 presents the results of 20 replications of this procedure. Recovery (hit rate) was good for most model versions, though poor for some. The best-fitting model of Experiment 5 had a hit rate of 1, but the best-fitting model of Experiment 6 had a hit rate of only .33; for the latter, the WAIC difference to the winning model was on average around 2. For comparison, the ΔWAIC in Experiments 5 and 6 were 5.5, and 5.1, respectively. False-alarm rates were negligible for all model versions, confirming that none of them has an undue advantage due to excess flexibility.

Parameter Recovery. We investigated parameter recovery for the model version that best fit Experiment 5. Again, we ran five recovery experiments, each varying one of the five parameters while keeping the other four at their best-fitting group-mean value from

[7] To make this computationally feasible we reduced the number of MCMC steps to 5000, preceded by 1000 burn-in steps.

Experiment 5. The parameters were varied across the following values (with the values closest to the posterior group means from Experiment 5 in bold):

$\mu(c)$ = [1, 1.5, **2**, 3, 4, 5, 7, 10],

$\mu(a)$ = [0.03, 0.05, **0.1**, 0.15, 0.2, 0.3, 0.5, 1],

$\mu(r)$ = [0.2, 0.3, **0.5**, 0.8, 1.2, 2, 3, 5],

$\mu(e)$ = [0.5, 1.0, 1.5, 2.0, 3.0, **4.0**, 6.0, 8.0],

$\mu(d)$ = [0.1, 0.2, 0.35, 0.5, **0.65**, 0.8, 0.9, 0.95].

Each simulation generated data for the design of Experiment 5, with N=30 and 100 trials for each of the CWI × WCI conditions.

The results are presented in Figures 29 to 33, each of which shows how the parameter estimates respond to manipulation of one parameter. The parameter estimates are less precise than for the complex-span model, but otherwise behave largely as expected: Each parameter manipulation systematically affected the manipulated parameter, and had at most a weak and less systematic effect on the others.

Taken together, the model recovery and parameter recovery simulations show that M^3 models are useful for two purposes: For determining which WM processes operate in an experimental setting and which dimension of memory strength (memory for elements or memory for bindings) they affect, and for measuring the theoretical variables reflected by the parameters. Model recovery is not perfect, but where it fails, it usually fails by a small margin on the WAIC scale. It follows that in reality, when two model versions differing with regard to an assumption (e.g., whether or not distractors are removed, or whether extended encoding affects only *a* or only *c*) differ by ΔWAIC of 10 or more when fit competitively to a given

data set, we can be fairly confident that the assumptions included in the winning model are more adequate for the experiment in question than those of the losing model.

Parameter recovery also is far from perfect, simply because the limited amount of information from an experiment of typical size does not enable highly precise parameter estimates. Nevertheless, the M^3 models passed two important tests of parameter recovery. First, there was no sign of systematic parameter trade-offs, and second, the precision of parameter estimates is reasonably well expressed in the breadth of the posterior distributions, as measured by their 95% HDIs. One reason why in the present M^3 applications the HDIs are fairly broad is that we started from highly uninformative priors. The Bayesian framework enables us to improve the precision of parameter estimates by accumulating evidence across experiments: The posteriors of parameter values from one experiment can be used as the priors for the next experiment with the same task and a similar design. To the degree that similar parameter values are credible for successive experiments, the precision of parameter estimates should increase (Kary, Taylor, & Donkin, 2016; Lee & Vanpaemel, 2018).

General Discussion

In this article we introduced the M^3 framework for building simple measurement models for working-memory tasks. We provided initial evidence for the core model components, memory for individual elements (a) and memory for content-context bindings (c), through a selective-influence experiment. Experiment 1 demonstrated that the parameters thought to capture the presumed latent psychological constructs selectively responded to a targeted experimental intervention. We then demonstrated the use of M^3 for analyzing experimental designs by applying measurement models to two variants of the complex-span paradigm and to the memory-updating paradigm. Moreover, we demonstrated the use of M^3 for analyzing individual and group differences by a study of age differences in model parameters across both paradigms. We will next discuss the M^3 framework in the context of

other modelling approaches, and then discuss the implications of the present findings for decomposing representations and processes in working memory, and their implications for age differences in working memory.

The Memory Measurement Models Framework

The M^3 framework is intended as a generic tool for building measurement models for experimental working-memory tasks. The framework can be applied to any task in which participants select their responses from a set of candidates that represent several response categories. The categories are assumed to differ in the degree of activation they receive at test from the information in memory in conjunction with the available retrieval cues. We designed the framework for working-memory tests, but in principle it could also be applied to tests of episodic long-term memory that meet the above requirements.

Compared to fully-developed computational models of (working) memory, models built in the M^3 framework are obviously much simplified. They are not intended to capture many details that have informed more elaborate models, such as the serial-position curve or the gradient of transposition errors (Lewandowsky & Farrell, 2008). This simplification is intended because there is a trade-off between the level of detail a model captures and the robustness of parameter estimates that it yields. This trade-off has been documented for the case of the drift-diffusion model of response-time distributions: Whereas the full model version explains the data in more detail than simpler model versions (Ratcliff & Rouder, 1998), simulations have shown that a much simplified version, called the EZ diffusion model (Wagenmakers et al., 2007), is better suited for recovering model parameters for individual-differences studies (van Ravenzwaaij & Oberauer, 2009) and provides more power for detecting experimental effects on parameters (van Ravenzwaaij, Donkin, & Vandekerckhove, 2017). Therefore, we expect that the simplicity of M^3 pays off in terms of robustness of parameter measurement.

M^3 are similar to multinomial process-tree (MPT) models of memory (Batchelder & Riefer, 1999; Buchner et al., 1995; Schweickert, 1993) in that they are applicable to multinomial data, but they differ from MPT models in describing memory representations as varying continuously in strength, whereas MPT models of memory rest on the assumption of discrete states of remembering (or not remembering) some piece of information. We do not take a strong stance on the question whether information retrieved from memory is best described as varying continuously in strength or as resulting in discrete states of remembering – the present experiments were not carried out with the aim to adjudicate between these alternatives. Our aim is merely to add a continuous-strength modeling framework to our toolbox of measurement models for multinomial memory data.

In our view, an advantage of continuous-strength models is that they are closer to the likely mechanisms of memory, whereas discrete-state models describe the outcome of retrieval. All detailed models of the representations and processes of memory – from MINERVA (Hintzman, 1986) to SAM (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981) to REM (Shiffrin & Steyvers, 1997) and ACT-R (Anderson & Lebiere, 1998) – assume representations varying on continuous dimensions of strength, activation, or degree of match to retrieval cues. The same is true for models of recall from working memory (e.g., Burgess & Hitch, 1999; Lewandowsky & Farrell, 2008; Oberauer, Lewandowsky, et al., 2012). It is therefore straightforward to let hypothetical processes such as extended encoding or removal modify parameters reflecting memory strength. It is less straightforward to let them modify parameters of discrete-state models that reflect the probability of entering a certain state.

The M^3 framework shares the assumption of continuous memory strength with signal-detection theory (SDT) models of recognition memory. Whereas most SDT models assume a single dimension of memory strength, M^3 incorporates the distinction of two dimensions: The strength of individual elements, and the strength of bindings between elements and their

contexts. These two dimensions bear close similarity to the two dimensions of memory strength in two-dimensional SDT models of recognition (Göthe & Oberauer, 2008; Rotello, Macmillan, & Reeder, 2004; Wixted & Mickes, 2010). In these models, the dimension of familiarity reflects the strength of a global match signal from memory in response to the recognition probe, which provides information about whether or not the probe has been experienced in the relevant context (e.g., in the current memory list). The dimension of recollection reflects the amount of contextual detail that can be retrieved about the experience matching the probe – for instance the color or location in which the probe had been presented as part of the memory list, or other list items that had been presented right before or after the probe. Familiarity could be interpreted as reflecting the strength of memory for individual elements, weighted by their degrees of match to the probe. Recollection arguably reflects the strength of memory for relations between memory elements matching the probe and their context. Hence, we can think of M^3 as transferring the core assumptions of two-dimensional SDT models to recall.

Varieties of Testing Memory

One could object at this point that asking participants to reconstruct a memory set from a given set of candidates, rather than to generate the responses, changes the nature of the task to a degree that it is no longer a recall task. At least for the domain of working memory, such a stance would be difficult to uphold: Recall tests of working memory routinely use digits or letters as materials, for which there is a naturally well-defined set of response candidates. Recalling a list of digits in their correct order is tantamount to selecting, at each output position, one out of nine digits. Spatial working memory is also often tested with procedures in which participants reconstruct the serial order of a limited set of locations, or reproduce a pattern by selecting cells in a grid (e.g., Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001; Parmentier & Andrés, 2006). In support of this argument, direct comparisons

of reconstruction with standard recall of letter lists showed analogous effects of experimental manipulations on both: Temporal isolation of items affects memory with free reconstruction and free recall, but not with serial recall and enforced serial reconstruction (Lewandowsky, Nimmo, & Brown, 2008).

The case of word recall is less obvious because there is no objectively defined set of candidates, but even recall of words from working memory can be described as selection from a set of candidates constructed by the person. Most theories and models of serial recall of words involve a redintegration step at which the initially retrieved memory trace is disambiguated by matching it against representations of words in the mental lexicon, choosing the word that best matches the retrieved trace (Farrell & Lewandowsky, 2002; Goh & Pisoni, 2003; Hulme, Newton, Cowan, Stuart, & Brown, 1999; Thorn, Gathercole, & Frankish, 2005). Redintegration is effectively selection of a recall option from a set of candidates according to the relative strength of evidence from memory that each of them receives. In support, a direct comparison of both methods also shows that two experimental variables – serial position and word frequency – affect serial-order reconstruction and serial recall of words in the same way (Quinlan, Roodenrys, & Miller, 2017)

Taking a broader perspective, the different labels we use to describe the multitude of procedures for testing (working) memory – such as "recall", "recognition", or "reconstruction" – do not map well onto the different decision processes we tap through these procedures. Many methods described by these labels share the requirement to select one element from a set of candidates. They can be arranged on a continuum of methods varying in the size of the candidate set, with two-alternative forced-choice (2-AFC) recognition on one end, and recall of words sampled without replacement from a large, "open" pool on the other end. In contrast, old-new recognition requires a decision on whether a single probe is or is not a member of the current memory set. M^3 can be applied to all methods requiring selection from a candidate set,

as long as the candidate set is known, either because it is naturally defined (as in the case of letters, digits, or grid cells) or because it is given by the experimenter. Therefore, we hope that the M^3 framework will prove useful in bridging the gap between research on 2-AFC recognition and on recall.

The M^3 framework also facilitates bridging between research on working memory for discrete stimuli such as digits, words, or spatial locations in a grid on the one hand, and working memory for continuously varying features such as color or orientation on the other hand. Features varying on a continuous dimension have been the material of choice in many studies of working memory for simple visual materials (Bays, Catalao, & Husain, 2009; Luck & Vogel, 1997; Wilken & Ma, 2004; Zhang & Luck, 2008). Visual working memory is often studied with the continuous-reproduction paradigm (a.k.a. delayed estimation), in which participants are asked to reproduce the feature of one object – selected at random from the current memory set – on a continuous scale (Blake, Cepeda, & Hiris, 1997; Wilken & Ma, 2004). For instance, participants might be asked to remember an array of several colored dots, and at test they select the color of one of the dots from a color wheel. One of us has developed a measurement model for this task, called the interference measurement model (IMM) (Oberauer et al., 2017). The IMM is a continuous-strength alternative to an earlier measurement model – the so-called mixture model – that builds on the notion of discrete memory states (Zhang & Luck, 2008). The IMM has much in common with the present M^3: The probability of selecting each color from the color wheel is proportional to the activation they receive at test. Activation of the color of each item in the array receives activation from two sources, one reflecting the strength of binding of the target to the location that serves as the retrieval cue, corresponding to M^3 parameter c, and one reflecting memory strength of all items regardless of their location in the array, corresponding to parameter a. Both strength parameters are measured relative to a baseline activation b assigned to all colors in the color

wheel. The main difference to M^3 for discrete stimuli is that in the IMM for continuous reproduction the response candidates – the colors on the color wheel – have a similarity structure, such that activation of each item's color spreads to neighboring colors on the color wheel according to the precision of feature representations.

Taken together, the M^3 framework completes the matrix of measurement-model frameworks for (working) memory (see Table 4), so that we now have discrete-state and continuous-strength modeling frameworks for the three most common forms of testing working memory: old-new recognition, selection from a set of discrete candidates (whether we call this multi-alternative recognition, reconstruction, or recall), and continuous reproduction.

Implications for Processes in Working Memory

The M^3 framework can be used to analyze experimental effects on the two core parameters, strength of memory for elements (potentially relying on persistent activation of representations) and strength of memory for relations (relying on content-context bindings). A simple way of doing so is to apply the basic M^3 to each experimental condition and compare the estimates of parameters *a* and *c*. For the demonstrations in this article we chose a more sophisticated approach in which the experimental effects are captured by additional parameters that reflect the hypothetical processes responsible for the effects. For instance, the difference between a memory item and a distractor in the complex-span paradigm is captured by the filtering parameter *f*. With this approach we can gauge the contributions of several hypothetical processes to an experimental effect. For instance, the beneficial effect of longer free time in the complex-span paradigm could be explained by removal of distractors or by extended encoding of memory items, or a combination of both. With the M^3 framework we can measure the two processes separately, capitalizing on the fact that they have different effects on the relative strength of response candidates: Distractor removal reduces the

memory strength of distractors relative to memory items and to NPLs, while the relative strength of memory items and NPLs remains constant. In contrast, extended encoding boosts the strength of memory items relative to distractors and NPLs while keeping the ratio of the latter two constant.

Extended Encoding. Across five experiments we found that free time is used for extended encoding, defined as gradual strengthening of memory representations over time after offset of stimulus presentation. Extended encoding can be interpreted as resulting from at least four processes proposed to play a role in working memory for verbal materials: articulatory rehearsal, refreshing, short-term consolidation, or elaborative rehearsal. As discussed in detail after Experiments 2 to 4, there are subtle differences between these hypothetical processes, and evidence speaking to their suitability for explaining the extended-encoding effect is sparse and ambiguous. Short-term consolidation is unlikely to provide a full explanation for extended-encoding effects because extended encoding also uses free time intervals not immediately following presentation of the strengthened memory items. Future experiments could be tailored to disentangle the contributions of the remaining three processes and measure their effect through the *e* parameter in appropriate M^3 versions. For instance, researchers could direct articulatory rehearsal to a subset of list items through instruction, or guide refreshing to a subset through refreshing cues (Souza et al., 2015), and test the prediction that extended encoding effects are limited to that subset. The contribution of elaboration could be gauged by comparing memory materials that are easy or hard to elaborate.

In decay models of working memory, articulatory rehearsal and refreshing have been assumed to counteract decay (Camos et al., 2009). A process that uses free time to restore previously decayed memory traces would be captured by the extended-encoding parameter in the M^3 for the present experiments – in that case the parameters *a* and *c* would not reflect the

strength achieved after initial encoding, but the level reached after a certain amount of decay. These experiments were not designed to measure decay separately from other processes, so they provide no information on whether or not decay occurs. Other experiments designed to answer this question have shown that verbal contents of working memory do not decay (Oberauer & Lewandowsky, 2013, 2014), and therefore we do not find this interpretation plausible, but the issue is still under debate (Ricker, Vergauwe, & Cowan, 2016). Proponents of decay models could use the *e* parameter to gauge the efficiency of memory restoration through articulatory rehearsal or refreshing. Future applications of M^3 to experimental designs suited to identify a hypothetical effect of decay could be used to determine whether this interpretation is tenable. Even if it is not, articulatory rehearsal could still contribute to extended encoding: Articulating the words could strengthen their memory representations by generating a phonological and an articulatory code in addition to the initial visual and/or semantic representation of the words. We are skeptical about this possibility for theoretical (Lewandowsky & Oberauer, 2015) and empirical reasons (Souza & Oberauer, 2018), but the matter certainly deserves further investigation.

Filtering and Removal. In the standard complex-span paradigm, distractors are known to be distractors before they are processed, so people can try to avoid encoding them into working memory. The degree to which they succeed in doing so is captured by the filtering parameter *f*. After a distractor has been encoded into working memory, it could be removed again. The notion of distractor removal is part of the SOB-CS model of complex span (Oberauer, Lewandowsky, et al., 2012). In SOB-CS, removal is conceptualized as a gradual process that uses free time following a distractor to remove it. In the M^3 this gradual removal is captured by parameter *r*. We included gradual removal also in the updating models because in the updating task old memory items need to be removed from working memory to minimize proactive interference. The evidence from the present experiments for a gradual

removal process as envisioned in SOB-CS is mixed. On the positive side, the best-fitting model always included the *r* parameter – in no case did a model version without any role for gradual removal win the competition (smallest ΔWAIC of a no-removal model compared to the winning model = 25). On the negative side, in the complex-span experiments, distractor removal was identified as a very rapid process, which is largely complete even after fairly short free-time intervals (see Oberauer, 2018, for further evidence that, under some conditions, removal is very rapid). Therefore, contrary to SOB-CS, removal is unlikely to contribute much to explaining the beneficial effect of free time in complex-span tasks. This effect is more likely due to extended encoding.

Rapid removal also must be assumed in the updating paradigm: Memory strength of old items was weakened relative to current items largely independent of free time. In the models we captured this time-independent removal through parameter *d*. As in complex span, this process affected only bindings of old items to their context (parameter *c*). It might be accompanied by a slow gradual removal process, captured by *r*, but whether it affected *a* or *c* and its estimated values, were inconsistent between the two updating experiments, so we cannot draw strong conclusions about it.

Taken together, the experiments reveal several processes involved in controlling the contents of working memory: (1) Filtering of distractors known to be distractors already during encoding. (2) At least one removal processes, proceeding very fast and affecting bindings. (3) Potentially a second, slower removal process helping to get rid of old items in memory updating.

Implications for Adult Age Differences in Working Memory

The M^3 for complex span and WM updating afford a decomposition of age differences in working memory: Older adults are selectively impaired in building and maintaining

bindings between content elements and their contexts. In contrast, their ability to maintain memory for individual elements encountered in the current trial – perhaps through persistent activation – was not at all impaired. If anything, older adults' *a* parameters were larger than those of younger adults. The selective age-related binding deficit converges with previous research that pointed towards an age-related binding deficit in working memory (Mitchell, Johnson, Raye, Mather, & D'Esposito, 2000; Oberauer, 2005; Peterson & Naveh-Benjamin, 2016) and in episodic long-term memory (Chalfonte & Johnson, 1996; Naveh-Benjamin, Hussain, Guez, & Bar-On, 2003; Old & Naveh-Benjamin, 2008). The finding that age-related differences in working memory capacity are due to differences in content-context bindings but not in the ability to maintain individual elements lends support to the binding hypothesis of individual differences in working-memory capacity (Oberauer, Süß, Wilhelm, & Sander, 2007): Differences in capacity arise from differences in the ability to create and maintain temporary bindings between elementary representations.

No other age differences were observed consistently across both tasks. For the updating task, but not the complex-span task, extended encoding – further boosting bindings – was reduced in old age. Filtering of distractors was equally effective in both age groups. Rapid removal of distractor-context bindings was also equally effective for young and old adults in the complex-span experiments. In contrast, older adults were somewhat less effective in rapidly removing old item-position bindings in the updating task (parameter *d*). These results lend at best partial support to the hypothesis that age-related impairments in working-memory functioning are due to impaired inhibition of irrelevant information (Hasher & Zacks, 1988). Hasher et al. (1999) have distinguished three aspects of inhibition, two of which are directly relevant for the control of the contents of working memory: control of *access*, and *deletion*. Control of access of information to working memory is measured by the filtering parameter in the M^3 of complex span, and we found no evidence that old adults are

impaired in this function. Deletion is measured by parameters *r* and *d* in the M^3; we found evidence for age differences only in *d*. In partial agreement with our results, one previous study found age differences in deletion of information from working memory, but not in the control of access (Cansino, Guzzon, Matrinelli, Barollo, & Casco, 2011).

The age-related difference in content-context bindings could be interpreted as resulting from an age-related difference in the ability to discriminate temporal contexts (Dumas & Hartman, 2003; Maylor, Vousden, & Brown, 1999) or the ability to reinstate the appropriate temporal context needed as retrieval cue for associated content (Healey & Kahana, 2016). This interpretation agrees well with the complex-span data, where older adults' reduced *c* parameter translates into an impaired ability to discriminate between items (and distractors) in the to-be-recalled list position and those in other list positions. The temporal-context explanation works less well for the memory-updating paradigm: Here, the reduced *c* parameter translates into older adults' impaired ability to discriminate between items in the probed position from items in other positions. In the updating task these positions differed along the spatial left-right dimension, which was not correlated with the temporal order of word presentation (except for the initial four words). Compared to their discrimination problem on the spatial dimension, older adults were considerably less impaired in discriminating between the current items and the old items in each position. If older adults had particular difficulties with discriminating events in time, we might expect them to suffer more strongly from confusions along the time dimension (i.e., proactive interference from old items). To conclude, the common pattern in both tasks is that older adults have difficulties discriminating between items within the current memory set, whereas they are only mildly impaired in discriminating between relevant information (i.e., the current memory set) and irrelevant information (i.e., distractors, or old items). This common pattern of age differences maps well onto the assumption – incorporated in M^3– that the discrimination of elements

within the current memory set relies on content-context bindings, whereas the discrimination between currently relevant and irrelevant material relies on processes that filter or remove the irrelevant material.

Comparison of Best-Supported Model Assumptions Across Experiments

The model versions that won the competition in each experiment have much in common, but there were two instances in which different models came out best: (1) The best-fitting model for the standard complex-span paradigm (Experiments 2 and 6) differed from the best-fitting model for the pre-cue/post-cue version of complex span (Experiments 3 and 4) in one regard: For standard complex span, extended encoding affected both memory for individual elements and memory for bindings, whereas for the pre-cue/post-cue design, extended encoding affected only bindings. (2) The best-fitting models for the two updating experiments (Experiments 5 and 6) differed in two regards: First, in Experiment 5, removal affected only memory for elements, whereas in Experiment 6, it affected only memory for bindings. As mentioned in the Discussion of Experiment 6, the removal parameter in that experiment was estimated to a value where removal had practically no effect, so this discrepancy is probably meaningless. Second, the best-fitting updating models again differed in whether extended encoding affected both memory for elements and memory for bindings (Experiment 5) or only memory for bindings (Experiment 6). Taken together, we see a remarkable degree of convergence between the assumptions in the best-fitting models across experiments, with one exception: Sometimes extended encoding affected memory for individual elements (in addition to memory for bindings), and sometimes it did not. Future research might investigate whether this is a true difference, and if so, which variables control it.

One next step in using and exploring the present framework is to apply M^3 to multiple experiments simultaneously. For instance, the M^3 for complex span and for updating could be

fit jointly to the data of Experiment 6, using the same model versions with regard to processes shared between the models (i.e., extended encoding and removal), and perhaps even the same values for their shared parameters. Within the hierarchical modelling framework, M^3 could also be applied jointly to experiments run with different participant samples: The parameters for each participant could still be drawn from the same distribution, governed by the same population-level parameters. From a statistical point of view, doing this merely adds one more level to the hierarchy: Each experiment is a sample from the population of possible experiments designed to measure a subset of the parameters that are included in (or can be added to) M^3. Jointly fitting multiple experiments therefore yields similar advantages to jointly fitting multiple participants in a hierarchical framework: Data from each experiment constrain the parameter estimates for other experiments. We can ask whether there are systematic differences between experiments in the same way as we can ask whether there are systematic differences between individuals (Thiele, Haaf, & Rouder, 2017). From a theoretical point of view, the additional advantage is that we can ask whether the latent variables and processes that we give the same interpretation in measurement models for different tasks (e.g., memory for bindings, removal) are actually the same: If they are, the parameters describing these variables or processes should be systematically related – in the simplest case, they should be the same.

Concluding Comment

Some readers might object that the conclusions from the present experiments are valid only if we accept the assumptions made in the measurement models. This is true—but it is true not only for the M^3 framework, but for all measurement models, whether they are explicitly formulated or implicitly assumed. Every time researchers make an inference from one or several observable variables to one or more latent variables they use a measurement model. Often this measurement model is left implicit, and researchers simply take for granted

that their dependent variable reflects the construct of interest. The implicit measurement model of most research is a simple monotonic mapping between one observed variable and the one latent variable it is meant to measure, tacitly assuming that the observed variable is a process-pure measurement of the intended latent variable. Often an additional tacit assumption is that the mapping is linear (Loftus et al., 2004). When the measurement model is left implicit, it is no less fallible than when it is made explicit, but with an implicit model we are unlikely to question it. Relative to the measurement model implicit in most research, using models built within the M^3 framework, or other comparable frameworks such as SDT or MPT models, has two advantages. First, the measurement model is made explicit and thereby exposed to critical scrutiny and empirical test. Second, these measurement models rest on assumptions that are at least plausible, whereas the implicit assumption of a one-to-one, linear mapping between one observed and one latent variable is often untenable upon close inspection (Jacoby, 1991; Loftus et al., 2004). Therefore, the way forward for those who are skeptical about the assumptions incorporated in M^3 is not to eschew explicit measurement models but to build better ones.

Appendix A: Justification of Modelling Decisions

In addition to core theoretical assumptions, every computational model incorporates a number of auxiliary assumptions that are necessary to make the model work, but which are not chosen for theoretical reasons. These assumptions might appear arbitrary, although often they are not. Here we make explicit the reasons for the auxiliary assumptions we had to make for the M^3 framework:

1. We used a version of Luce's choice rule to translate the activation of each response candidate into the probability of choosing it. This decision rule is not without alternatives. We ran simulation studies comparing our version of Luce's choice rule – which normalizes activation *A* directly – to a set of alternatives that have been discussed in the literature on choice and decision making: (a) Another version of Luce's choice rule, which normalizes exp(*A*) instead of *A*; this version has been proven to be equivalent to an n-alternative SDT model with a Gumbel (or "double-exponential") noise distribution (Yellott, 1977); (b) an n-alternative SDT model with Gaussian noise (DeCarlo, 2012), (c) the Linear Ballistic Accumulator (LBA) model, (S. D. Brown & Heathcote, 2008), and (d) the Leaky Competing Accumulator (LCA) model (Usher & McClelland, 2001). The latter two are models of multi-alternative response times and the associated response probabilities; here we considered only their predictions about response probabilities. We simulated predictions of the M^3 for complex span (Experiment 2) combined with these five decision rules, varying the 5 parameters (*a*, *c*, *f*, *r*, and *e*). We found that the rules fall into two clusters that differ qualitatively in how the pattern of predicted response probabilities reacts to these parameter changes. One cluster was formed by the simple version of Luce's choice rule we built into M^3 together with the LBA and the LCA; the other cluster was formed by Luce's choice rule on exp(A) and the Gaussian SDT (see Figure A1).

What distinguishes the two clusters of decision rules is that the first cluster builds on the assumption that activation values A are expressed on a ratio scale (Luce, 1977) with a true zero point that means "no activation", whereas the second cluster treats these values as lying on an interval scale on which zero is an arbitrarily chosen point. This implies that, in the first cluster of rules, multiplying all A values by a constant leaves the choice probabilities unchanged – this is obvious for Luce's choice rule, and approximately true also for LBA and LCA. In the second cluster, adding a constant to all A values leaves the choice probabilities unchanged. For instance, in SDT models the probability of choosing one response option over one or several others depends on d', which is the difference between the signals associated with the response choices. Adding a constant to all signals does not affect d' and therefore does not change the choice probabilities.

We decided to use the version of Luce's choice rule that normalizes *A* directly for two reasons. First, we found a ratio scale for activation values desirable, and Luce's choice rule is the simplest way of implementing it. On a ratio scale an activation value of zero is meaningful – it means that a response candidate has no evidence from memory in its favor, so that it can be ruled out with certainty. This would be the case, for instance, if we asked a person to remember a list of words and then presented them with a set of response candidates including numbers and pictures. We expect that people never choose any number or picture based on normal memory processes, and we want models in the M^3 framework to be able to assign such choices zero likelihood. A ratio scale also facilitates interpretation of the filtering parameter as a proportional reduction of memory strength, and of the removal parameter as reducing memory strength by a certain proportion over a certain amount of time. Second, by using a decision rule from the first cluster, we designed the M^3 framework in such a way that extending it to response time distributions does not entail a qualitative change in model behavior, because LBA and LCA fall into the same cluster. In particular, we envisage that

combining M^3 with LBA, which was designed with the aim to be the "simplest complete model of choice response time" (S. D. Brown & Heathcote, 2008), extends the models' explanatory range with minimal additional machinery. For these reasons we argue in favor of a decision rule that places activation values on a ratio scale, and we propose a version of Luce's choice rule as the simplest instantiation of such a decision rule for now; later extensions of M^3 could replace it with an evidence-accumulation decision process such as LBA or LCA.

2. A second auxiliary assumption is that memory strengths for elements and for bindings are combined additively. An equally parsimonious alternative is to combine them multiplicatively. This would make them non-compensatory: If either memory for elements or for bindings were zero, the activation value for that response option would be zero regardless of the other source of memory strength. Therefore, the model would predict that list items other than the correct item (response category "other") from very different contexts, as well as distractors that are bound to very different contexts than the currently tested item, are recalled only negligibly more often than not-presented lures. This is not the case: In a previous project investigating visual working memory we found that there was a tendency to recall non-targets even when they were far away from the target location in the array, which could be captured only by an additive parameter *a* in the model (Oberauer & Lin, 2017). In the present experiments, the tendency to recall distractors more than not-presented lures was observed even for distractors in serial positions very distant from the target (see Figure A2 for data from the complex-span tasks in Experiments 2 and 6). Therefore, combining the two strength parameters *a* and *c* additively seems more appropriate than multiplying them. An additive combination of these parameters also follows the precedent of 2-dimensional SDT models of recognition, in which familiarity and recollection are added into a signal (Göthe & Oberauer, 2008; Wixted & Mickes, 2010).

3. We made assumptions about the functional form of the time course of extended encoding and removal: Extended encoding increases memory strength linearly over time; removal reduces it exponentially towards zero. There are of course many alternatives to these functions, but as all experiments in the present manuscript manipulated time only over two levels, it would be impossible to empirically distinguish them: Any possible function makes equivalent predictions. This implies that our conclusions from the present experiments do not depend on these assumptions. Future work varying time in a more graded fashion will help to determine the best functional form for extended encoding and removal – with the constraint that the function for removal should asymptote towards zero. Hence, in contrast to the first two auxiliary assumptions discussed here, we see this third set as merely a choice of convenience that is not meant to stay.

Appendix B: Methods and Descriptive Results of Experiments

Experiment 1: Selective-Influence Test

Participants. Forty students of the University of Zurich took part in a one-hour session for course credit or reimbursement of 15 CHF (about 15 USD).

Design and Materials. The stimuli were 226 German nouns referring to concrete objects. Each trial involved 15 words chosen at random without replacement from the word pool; these words constituted the candidate set for recall for that trial. Five of them served as memory items; in the control condition, an additional five served as distractors. The remaining words were not-presented lures that appeared only in the test array.

Participants completed 20 trials for each of the three conditions, in random order, preceded by three practice trials, one from each condition, in random order.

Procedure. Each trial began with a central fixation cross, replaced 1 s later by the first memory word in red against a white background. Each memory word was followed by a distractor in black, then the next memory word, and so on until the fifth distractor. Each word was displayed for 0.9 s, followed by a blank screen for 0.1 s. Participants were instructed to read each word (both red and black) aloud, and to remember only the red words. After offset of the last word, the 15 recall candidates were displayed in a 5 x 3 array; each word was surrounded by a thin black frame. The words were assigned to their locations at random. Participants clicked with the mouse on the memoranda in order of presentation. Candidates already selected stayed on the screen unchanged and could be selected again. We did not eliminate already-chosen candidates in this or any of the following experiments because keeping the candidate set constant makes applying the M^3 framework easier. Once

participants had selected as many words as there were memoranda, the candidate set was replaced by display of the message “Continue by pressing the space bar”.

Experiments 2, 3, and 4: Complex Span Tasks

Participants. Participants in Experiment 2 were 27 young adults from the University of Western Australia community who took part in exchange for partial course credit. Experiments 3 and 4 enrolled students from University of Zurich as participants (N = 24 and 26, respectively). They received either partial course credit or 15 CHF in compensation for a one-hour session. Two participants in Experiment 4 did not respond to a single size-judgment task in time in at least one condition, and therefore were removed from all analyses, resulting in N=24 for that experiment.

Design and Materials. Participants started the experiment with three practice trials from conditions sampled at random. In Experiment 2, they went on to complete 40 test trials, 20 from each condition (short vs. long free time), mixed at random. In Experiments 3 and 4, they completed four blocks of 15 trials each. Conditions varied between blocks in an order that was counterbalanced across participants.

The stimuli consisted of 506 English nouns (Experiment 2) or 226 German nouns (Experiments 3 and 4) referring to concrete objects. Participants’ task was to judge for each word whether it represented an object larger or smaller than a soccer ball. Therefore, we assigned an estimate of the object’s size to each word based on our own judgment (1 rater). Both our size judgments and those of participants are to some degree subjective (e.g., whether a penguin is larger or smaller than a soccer ball depends on what kind of penguin you think of). This is not a matter of concern because we were not interested in accuracy of the size

judgments (as long as it was above chance); these judgments only served to ensure that participants processed all words.

Each trial involved 15 words chosen at random without replacement from the word pool. In Experiments 3 and 4, the words were sampled such that no word was used more than once within a block of 15 trials. The 15 words sampled for a trial constituted the candidate set for recall for that trial. Of the 15 candidates, ten were presented as memory items or distractors, and the remaining five served as not-presented lures (NPL) in the candidate set. Of the ten presented words, a subset was designated as memoranda and the remaining subset as distractors. In Experiment 2, there were always five memoranda and five distractors, presented in alternation. In Experiments 3 and 4 the test trials – used for analysis – also always had five memoranda and five distractors, presented in a random order. In addition, in these experiments there were 3 practice trials and 2 filler trials per condition, for which the number of memory items was chosen at random from a uniform distribution from four to seven, and the remaining words out of 10 were presented as distractors. This variation in the number of memoranda and distractors served to discourage participants from counting the number of memoranda and thereby anticipating the status of the last word or words in a trial.

Procedure Experiment 2. Each trial began with a central fixation cross, replaced 1 s later by the first memory word in red against a white background. Participants made their size judgment by pressing the left arrow key for “smaller” and the right arrow key for “larger”. The word stayed on the screen until participants responded or until 1.7 s had elapsed; this time allowance was based on prior experience with the size-judgment task in our lab, and imposes a moderate amount of time pressure. Each item was followed by a distractor, presented in black, on which participants again made a size judgment. Each distractor was followed by a free-time interval, during which the screen was blank. Depending on the condition of a trial, the five free-time intervals after each distractor were all 0.2 s (short) or 1.7 s (long). The

interval following a memory item was always 0.2 s. Twenty short and twenty long free-time trials were mixed in a random order. Memory was tested in the same way as in Experiment 1.

Procedure Experiments 3 and 4. Each trial began with a central fixation cross, followed 0.5 s later by the central presentation of the first word within a thick rectangular frame. Memory items and distractors were presented in a new random order in each trial. In the pre-cued condition, the frame was either blue, indicating a memory item, or red, indicating a distractor. Participants made a size judgment as in Experiment 2. In the post-cue condition, the frame was grey, and turned red or blue once the participant has made the size judgment; at the same time the word was erased. In both conditions, the frame remained visible for 0.5 s after the word disappeared. In the short free-time condition, offset of the frame was followed immediately by onset of the next word (and its frame). In the long free-time condition of Experiment 3, each distractor was followed by 1.5 s between frame offset and onset of the next word; during this interval the screen went blank. In Experiment 4, this free-time interval instead followed each item; in all other regards the two experiments were identical. After the last size judgment of a trial, memory was tested in the same way as in Experiment 1.

Experiments 3 and 4 each consisted of four blocks, one for each combination of time of cueing (pre vs. post) and free time (short vs. long). Order of blocks was counterbalanced across participants. Each block started with three practice trials, followed by ten test trials and two filler trials. The filler trials were placed at random in positions 6-8 (first filler) and 11-13 (second filler) within the 15-trial sequence of a block. Practice and filler trials were not included in the analyses.

Descriptive Results. Performance in the size-judgment task is summarized in Table A1. Across all three experiments, participants failed to respond before the deadline in less than 10% of trials. When they responded, their error rate – as assessed by comparison to our own judgments – was consistently below chance, showing that they took the size-judgment task

seriously. Performance in the size-judgment tasks did not vary conspicuously across experimental conditions.

The variable of primary interest was the number of responses in each of five categories: correct responses (i.e., selection of the correct item in its correct ordinal position), other items (i.e., memoranda from other positions), distractors from the to-be-recalled position (i.e., immediately preceding or following the item that would have been correct), distractors in other positions, and not-presented lures (NPL). These frequencies, summed across participants for each experimental condition, are presented in Tables A2 and A3.

We defined as distractors from the to-be-recalled position (short: *distractors in position*) those distractors that immediately preceded the to-be-recalled item in a given output position, or immediately followed it. In Experiment 2, in which memory items and distractors alternated regularly, there were two of these for all list positions except the first. In Experiments 3 and 4, items and distractors occurred in random order, so that the item to be recalled in a given output position was immediately preceded by a distractor in only about half of all trials, and likewise, it was immediately followed by a distractor in only about half of all trials. Therefore, the number of distractors in the candidate set that counted as "in position" was approximately one out of four distractors, but it varied randomly across conditions and participants. We computed the probability of selecting a *distractor in position*, plotted in the figures in the main text, by dividing the number of selected *distractors in position* by the number of *distractors in position* in the candidate set, separately for each participant and condition.

Experiment 5: Memory Updating Task

Participants. Twenty students of the University of Zurich took part in a single session lasting about one hour.

Materials and Procedure. The materials consisted of 368 German nouns referring to concrete objects. For each trial, four initial memory words were selected at random without replacement. Additional words were sampled for a variable number of updating step (between 0 and 20 steps), and to serve as four not-presented lures in the test array. Each word was used only once in each run of four consecutive trials.

Each trial began with the presentation of a row of four rectangular frames in the upper half of the screen. Simultaneously with the frames, a fixation cross appeared in the first (left-most) frame. This cue was presented for 200 ms. Depending on the cue-to-word interval condition, 100 or 1100 ms after the offset of the cue, the first memory word was presented in the first box for 500 ms. The next word was presented in the next frame in the same way, starting with the fixation cross 100 or 1000 ms after offset of the preceding word, depending on the word-to-cue interval condition. After the fourth initial memory word was presented, the sequence continued in the same way presenting the updating words. Whereas the four initial memory words were always presented from left to right, the updating words were presented in a random order across the four frames, with the constraint that every consecutive set of four updating words covered all four frames. Participants were instructed to always remember the last word they had seen in each frame.

Before each updating word the computer determined at random whether to end the trial or to continue with another updating word. The probability of ending the trial was set to 0.1 at each step, so that participants should build a constant expectation of being tested for the current memory set at any time during updating; however, a trial ended after a maximum of 20 updating steps. At the end of each trial, 12 words were presented as recall candidates: The last four words presented in the four boxes (i.e., the words participants should remember for the four boxes at this point in time), the next-to-last word in each box, and four not-presented lures. In trials with less than 4 updating steps, the non-existing next-to-last words were

replaced by additional not-presented lures; these trials were excluded from all analyses. The 12 recall candidates were arranged at random in a 3 × 4 matrix underneath the row of frames. Memory for the words in all four frames was tested in a random order by presenting a question mark in one of the frames. Participants were asked to select the word they remembered seeing last in that frame with the mouse.

The experiment was organized into four blocks, one for each condition (crossing cue-to-word interval and word-to-cue interval). Order of conditions was counterbalanced across participants. Each block continued until 8 valid trials – excluding trials with less than 4 updating steps – had been completed. Before the four test blocks, participants worked through three practice trials with cue-word-intervals and word-cue-intervals of 800 ms.

Descriptive Results. Responses were classified as correct, other current word, old word in probed position, other old word, and NPL. The frequencies of these response categories in each condition, averaged across participants, are given in Table A4.

Experiment 6: Age Differences in Complex Span and Memory Updating

Participants. The old adults are a subset of N=59 of the participants in a previous study who agreed to take part in the present study. From the same previous study, 25 young adults volunteered to take part in the present study; to reach a sample size at least as large as that of old adults, we recruited additional young people from the same population (i.e., students of University of Zurich) for a total N = 68. Participants in the previous study had been assessed with the Mini-Mental State (MMS; Folstein, Folstein, & McHugh, 1975) and the CERAD-Plus test battery (Satzger et al., 2001) to assess their overall cognitive status, and the SF-36 questionnaire (Bullinger, 1998, p. -36) to assess their health status. Data from these assessments for the present sample of old adults and the subsample of 25 young adults are summarized in Table A5, together with basic demographic data for all participants. We did

not make these assessments for the newly recruited young adults because they were sampled from the same population as the young adults from the previous study, and there were no exclusion criteria for young adults (for old adults, the exclusion criterion of an MMS score < 27 had already been applied in the previous study).

For the complex-span task, data from 8 old and 2 young adults were excluded because they had > 50% time-outs on the size-judgment task (0 young, 6 old), or committed more than 30% size-judgment errors (2 young, 2 old), raising doubts whether they have seriously attempted the size judgments.

Materials and Procedure. Participants took part in two sessions of approximately 90 minutes each. In each session they worked on one of the working-memory tasks, and a set of cognitive inhibition tasks for an unrelated study. Because the focus of this study was on individual differences rather than differences between tasks, the order of task administration was constant for all participants: They all worked on the memory-updating task in the first session and the complex-span task in the second session.

The complex-span task was exactly as in Experiment 2, except that the materials consisted of German rather than English nouns, and the maximum presentation time for words – during which a size judgment had to be made – was increased from 1.7 to 2.2 s to accommodate the slower processing speed of old adults. The memory-updating task was exactly as in Experiment 5, except that 13 valid trials (rather than 8) were completed in each condition, and the presentation time for words was increased from 0.5 to 1.2 s to accommodate old adults' slower speed. The time adjustments were made for both age groups.

Descriptive Results. Performance on the size-judgments of the complex-span task is summarized in Table A1. There are no noticeable differences between the two age groups in

size-judgment time-out rate or accuracy, but older adults took somewhat more time for these judgments.

Average frequencies of response categories, calculated as for the previous experiments, are presented in Table A2 for the complex-span task, and in Table A5 for the memory-updating task.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, N. J.: Erlbaum.

Awh, E., & Vogel, E. K. (2008). The bouncer in the brain. *Nature Neuroscience, 11*, 5-6.

Bakeman, R., & McArthur, D. (1996). Picturing repeated measures: Comments on Loftus, Morrison, and others. *Behavioral Research Methods, Instruments, & Computers, 28*, 584-589.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General, 133*, 83-100.

Bartsch, L. M., Singmann, H., & Oberauer, K. (2018). The effects of refreshing and elaboration on working memory performance, and their contributions to long-term memory formation. *Memory & Cognition*. doi:10.3758/s13421-018-0805-9

Batchelder, W. H., & Alexander, G. E. (2013). Discrete-state models: Comment on Pazzaglia, Dube, and Rotello (2013). *Psychological Bulletin, 139*, 1204-1212. doi:10.1037/a0033894

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57-86.

Bayliss, D. M., Bogdanovs, J., & Jarrold, C. (2015). Consolidating working memory: Distinguishing the effects of consolidation, rehearsal and attentional refreshing in a working memory span task. *Journal of Memory and Language, 81*, 34-50. doi:10.1016/j.jml.2014.12.004

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9*, 1-11.

Blake, R., Cepeda, N. J., & Hiris, E. (1997). Memory for visual motion. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 353-369.

Bopp, K. L., & Verhaeghen, P. (2005). Aging and verbal memory span: a meta-analysis. *Journal of Gerontology: Psychological Science, 60B*, P223-P233.

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear - or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 587-606.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*, 539-576.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive Psychology, 57*, 153-178.

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Towards unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General, 124*, 137-160.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review, 106*, 551-581.

Burgess, N., & Hitch, G. J. (2006). A revised model of short-term memory and long-term learning of verbal sequences. *Journal of Memory and Language, 55*, 627-652.

Camos, V., Lagner, P., & Barrouillet, P. (2009). Two maintenance mechanisms of verbal information in working memory. *Journal of Memory and Language, 61*, 457-469.

Cansino, S., Guzzon, D., Matrinelli, M., Barollo, M., & Casco, C. (2011). Effects of aging on interference control in selective attention and working memory. *Memory & Cognition, 39*, 1409-1422.

Chalfonte, B. L., & Johnson, M. K. (1996). Feature memory and binding in young and older adults. *Memory & Cognition, 24*, 403-416.

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences, 7*, 547-552.

Cowan, N., Blume, C. L., & Saults, J. S. (2013). Attention to attributes and objects in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 731-747. doi:10.1037/a0029687

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268-294.

Craik, F. I. M., & Watkins, M. J. (1973). The role of rehearsal in short-term memory. *Journal of Verbal Learning & Verbal Behavior, 12*, 599-607.

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences, 7*, 415-423.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450-466.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigation of recency effects. *Psychological Review, 112*, 3-42.

DeCarlo, L. T. (2012). On a signal detection approach to m-alternative forced choice with bias, with maximum likelihood and Bayesian approaches to estimation. *Journal of Mathematical Psychology, 56*, 196-207. doi:10.1016/j.jmp.2012.02.00

Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 130-151.

Dumas, J. A., & Hartman, M. (2003). Adult age differences in temporal and item memory. *Psychology and Aging, 18*, 573-586. doi:10.1037/0882-7974.18.3.573

Ecker, U. K. H., Lewandowsky, S., & Oberauer, K. (2014). Removal of information from working memory: A specific updating process. *Journal of Memory and Language, 74*, 77-90. doi:10.1016/j.jml.2013.09.003

Ecker, U. K. H., Oberauer, K., & Lewandowsky, S. (2014). Working memory updating involves item-specific removal. *Journal of Memory & Language, 74*, 1-15. doi:10.1016/j.jml.2014.03.006

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review, 119*, 223-271.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review, 9*, 59-79.

Farrell, S., & Lewandowsky, S. (2004). Modelling transposition latencies: Constraints for theories of serial order memory. *Journal of Memory and Language, 51*, 115-135.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing, 24*, 997-1016. doi:10.1007/s11222-013-9416-2

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 1*(457-511).

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1-67.

Goh, W. D., & Pisoni, D. B. (2003). Effects of lexical competition on immediate memory span for spoken words. *Quarterly Journal of Experimental Psychology, 56A*, 929-954.

Göthe, K., & Oberauer, K. (2008). The integration of familiarity and recollection information in short-term recognition: Modeling speed-accuracy trade-off functions. *Psychological Research, 72*, 289-303.

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implication for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 846-858.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 22, pp. 193-225). New York: Academic Press.

Hasher, L., Zacks, R. T., & May, C. P. (1999). Inhibitory control, circadian arousal, and age. In D. Gopher & A. Koriat (Eds.), *Attention and Performance* (pp. 653-675). Cambridge, MA: MIT Press.

Healey, M. K., & Kahana, M. J. (2016). A four-component model of age-related memory change. *Psychological Review, 123*(23-69). doi:10.1037/rev0000015

Heathcote, A., Brown, S. D., & Wagenmakers, E. J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann & E. J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 25-48). New York: Springer.

Henson, R. N. A. (1998). Short-term memory for serial order: The Start-End Model. *Cognitive Psychology, 36*, 73-137.

Hertel, P. T., & Calcaterra, G. (2005). Intentional forgetting benefits from thought substitution. *Psychonomic Bulletin & Review, 12*, 484-489.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411-428.

Houghton, G. (1990). The problem of serial order: A neural network model of sequence learning and recall. In R. Dale, C. Mellish, & M. Zock (Eds.), *Current research in natural language generation* (pp. 287-319). London: Academic Press.

Hulme, C., Newton, P., Cowan, N., Stuart, G., & Brown, G. (1999). Think before you speak: Pauses, memory search, and trace redintegration processes in verbal memory span. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 447-463.

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin, 140*, 339-373. doi:10.1037/a0034221

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513-541.

Johnson, M. K. (1992). MEM: Mechanisms of recollection. *Journal of Cognitive Neuroscience, 4*, 268-280.

Jolicoeur, P., & Dell'Acqua, R. (1998). The demonstration of short-term consolidation. *Cognitive Psychology, 36*, 138-202.

Kary, A., Taylor, R., & Donkin, C. (2016). Using Bayes factors to test the predictions of models: A case study in visual working memory. *Journal of Mathematical Psychology, 72*, 210-219. doi:10.1016/j.jmp.2015.07.002

Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1795-1804. doi:10.1037/xlm0000016

Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: a critical test with minimal assumptions. *Psychological Review, 122*, 542-557. doi:10.1037/a0039251

Kessler, Y., & Meiran, N. (2006). All updateable objects in working memory are updated whenever any of them are modified: Evidence from the memory updating paradigm. *Journal of Experimental Psychology: Learning, Memory & Cognition, 32*, 570-585.

Kessler, Y., & Meiran, N. (2008). Two dissociable updating processes in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1339-1348.

Kessler, Y., & Oberauer, K. (2014). Working memory updating latency reflects the cost of switching between maintenance and updating modes of operation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 738-754.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York: Academic Press.

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review, 25*, 114-127. doi:10.3758/s13423-017-1238-3

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian modeling for cognitive science: A practical course*. Cambridge: Cambridge University Press.

Lewandowsky, S., & Farrell, S. (2008). Short-term memory: new data and a model. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 49, pp. 1-48). London, UK: Elsevier.

Lewandowsky, S., Nimmo, L. M., & Brown, G. D. A. (2008). When temporal isolation benefits memory for serial order. *Journal of Memory and Language, 58*, 415-428.

Lewandowsky, S., & Oberauer, K. (2015). Rehearsal in serial recall: An unworkable solution to the non-existent problem of decay. *Psychological Review, 122*, 674-699. doi:10.1037/a0039684

Lewis-Peacock, J. A., Kessler, Y., & Oberauer, K. (2018). The removal of information from working memory. *Annals of the New York Academy of Science, 1424*, 33-44. doi:10.1111/nyas.13714

Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review, 111*, 835-863.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology, 15*, 215-233.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279-281.

Marshuetz, C. (2005). Order information in working memory: An integrative review of evidence from brain and behavior. *Psychological Bulletin, 131*, 323-339.

Maylor, E. A., Vousden, J. I., & Brown, G. D. A. (1999). Adult age differences in short-term memory for serial order: Data and a model. *Psychology & Aging, 14*, 572-594.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Mitchell, K. J., Johnson, M. K., Raye, C. L., Mather, M., & D'Esposito, M. (2000). Aging and reflective processes of working memory: Binding and test load deficits. *Psychology and Aging, 2000*, 527-541.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related: A latent-variable analysis. *Journal of Experimental Psychology: General, 130*, 621-640.

Naveh-Benjamin, M., Hussain, Z., Guez, J., & Bar-On, M. (2003). Adult age differences in episodic memory: Further support for an associative-deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 826-837.

Nieuwenstein, M., & Wyble, B. (2014). Beyond a mask and against the bottleneck: Retroactive dual-task interference during working memory consolidation of a masked visual target. *Journal of Experimental Psychology: General, 143*, 1409-1427. doi:10.1037/a0035257

Nishiyama, R., & Ukita, J. (2013). Articulatory rehearsal is more than refreshing memory traces. *Experimental Psychology, 60*, 131-139. doi:10.1027/1618-3169/a000179

Oberauer, K. (2003). Selective attention to elements in working memory. *Experimental Psychology, 50*(4), 257-269.

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology-General, 134*(3), 368-387.

Oberauer, K. (2018). Removal of irrelevant information from working memory: sometimes fast, sometimes slow, and sometimes not at all. *Annals of the New York Academy of Science, 1424*, 239–255. doi:10.1111/nyas.13603

Oberauer, K., Farrell, S., Jarrold, C., Pasiecznik, K., & Greaves, M. (2012). Interference between maintenance and processing in working memory: The effect of item-distractor similarity in complex span *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 665-685. doi:10.1037/a0026337

Oberauer, K., & Lewandowsky, S. (2011). Modeling working memory: a computational implementation of the Time-Based Resource-Sharing theory. *Psychonomic Bulletin & Review, 18*(1), 10-45.

Oberauer, K., & Lewandowsky, S. (2013). Evidence against decay in verbal working memory. *Journal of Experimental Psychology: General, 142*, 380-411.

Oberauer, K., & Lewandowsky, S. (2014). Further evidence against decay in working memory. *Journal of Memory and Language, 73*, 15-30. doi:10.1016/j.jml.2014.02.003

Oberauer, K., & Lewandowsky, S. (2016). Control of information in working memory: Encoding and removal of distractors in the complex-span paradigm. *Cognition, 156*, 106-128. doi:10.1016/j.cognition.2016.08.007

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review, 19*, 779-819. doi:10.3758/s13423-012-0272-4

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review, 124*, 21-59.

Oberauer, K., Stoneking, C., Wabersich, D., & Lin, H.-Y. (2017). Hierarchical Bayesian measurement models for continuous reproduction of visual features from working memory. *Journal of Vision, 17*, 1-27. doi:10.1167/17.5.11

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity - facets of a cognitive ability construct. *Personality and Individual Differences, 29*, 1017-1045.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 49-75). New York: Oxford University Press.

Oberauer, K., & Vockenberg, K. (2009). Updating of working memory: Lingering bindings. *Quarterly Journal of Experimental Psychology, 62*(5), 967-987.

Old, S. R., & Naveh-Benjamin, M. (2008). Differential effects of age on item and associative measures of memory: a meta-analysis. *Psychology and Aging, 23*, 104-118. doi:10.1037/0882-7974.23.1.104

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review, 105*, 761-781.

Parmentier, F. B. R., & Andrés, P. (2006). The impact of path crossings on visuo-spatial serial memory: Encoding or rehearsal effect? *Quarterly Journal of Experimental Psychology, 59*, 1867-1874.

Peterson, D. J., & Naveh-Benjamin, M. (2016). The role of aging in intra-item and item-context binding processes in visual working memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 42*, 1713-1730. doi:10.1037/xlm0000275

Plummer, M. (2016). JAGS 4.2.0. Retrieved from http://mcmc-jags.sourceforge.net/

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News, 6*, 7-11.

Quinlan, P. T., Roodenrys, S., & Miller, L. M. (2017). Serial reconstruction of order and serial recall in verbal short-term memory. *Memory & Cognition, 45*, 1126-1143. doi:10.3758/s13421-017-0719-y

R_Core_Team. (2017). R: A language and environment for statistical computing (Version 3.3.3). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from URL: http://www.R-project.org

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93-134. doi:10.1037/0033-295X.88.2.93

Rac-Lubashevsky, R., & Kessler, Y. (2016). Dissociating working memory updating and automatic updating: The reference-back paradigm. *Journal of Experimental Psychology: Learning, Memory and Cognition, 42*, 951-969. doi:10.1037/xlm0000219

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*, 347-356.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*, 438-481.

Raye, C. L., Johnson, M. K., Mitchell, K. J., Greene, E. J., & Johnson, M. R. (2007). Refreshing: A minimal executive function. *Cortex, 43*, 135-145.

Ricker, T. J., & Cowan, N. (2014). Differences between presentation methods in working memory procedures: a matter of working memory consolidation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 417-428. doi:10.1037/a0034301

Ricker, T. J., Vergauwe, E., & Cowan, N. (2016). Decay theory of immediate memory: From Brown (1958) to today (2014). *Quarterly Journal of Experimental Psychology, 69*, 1969-1995. doi:10.1080/17470218.2014.914546

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remembering and knowing: A two-dimensional signal-detection model. *Psychological Review, 111*, 588-616.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences, 105*, 5975-5979.

Salthouse, T. A. (1994). The aging of working memory. *Neuropsychology, 8*, 535-543.

Salthouse, T. A., Babcock, R. L., & Shaw, R. J. (1991). Effects of adult age on structural and operational capacities in working memory. *Psychology and Aging, 6*, 118-127.

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition, 21*, 168-173.

Sederberg, P. B., Howard, M. C., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review, 115*, 893-912.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM - retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145-166.

Souza, A. S., & Oberauer, K. (2018). Does articulatory rehearsal help immediate serial recall? *Cognitive Psychology, 107*, 1-21. doi:10.1016/j.cogpsych.2018.09.002

Souza, A. S., Rerko, L., & Oberauer, K. (2015). Refreshing memory traces: thinking of an item improves retrieval from visual working memory. *Annals of the New York Academy of Sciences, 1339*, 20-31. doi:10.1111/nyas.12603

Souza, A. S., Vergauwe, E., & Oberauer, K. (2018). Where to attend next: guiding refreshing of visual, spatial, and verbal representations in working memory. *Annals of the New York Academy of Science*. doi:10.1111/nyas.13621

Su, Y.-S. (2015). Package "R2jags" (Version 0.5-7).

Tan, L., & Ward, G. (2008). Rehearsal in immediate serial recall. *Psychonomic Bulletin & Review, 15*, 535-542.

Thiele, J. E., Haaf, J. M., & Rouder, J. N. (2017). Is there variation across individuals in processing? Bayesian analysis for systems factorial technology. *Journal of Mathematical Psychology, 81*, 40-54. doi:10.1016/j.jmp.2017.09.002

Thorn, A. S. C., Gathercole, S. E., & Frankish, C. (2005). Redintegration and the benefits of long-term knowledge in verbal short-term memory: An evaluation of Schweickert's (1993) multinomial processing tree model. *Cognitive Psychology, 50*, 133-158.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*, 550-592.

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2017). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychonomic Bulletin & Review, 24*, 547-556. doi:10.3758/s13423-016-1081-y

van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology, 53*(6), 463-473.

Vergauwe, E., Hardman, K. O., Rouder, J. N., Roemer, E., McAllaster, S., & Cowan, N. (2016). Searching for serial refreshing in working memory: Using response times to track the content of the focus of attention over time. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-016-1038-1

Vergauwe, E., Langerock, N., & Cowan, N. (2018). Evidence for spontaneous serial refreshing in verbal working memory? *Psychonomic Bulletin & Review, 2*, 674-680. doi:10.3758/s13423-017-1387-4

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 14*, 3-22.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and Widely Applicable Information Criterion in singular learning theory. *Journal of Machine Learning Research, 11*, 3571-3594.

Wei, Z., Wang, X.-J., & Wang, D.-H. (2012). From distributed resources to limited slots in multiple-item working memory: A spiking network model with normalization. *Journal of Neuroscience*(32).

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *frontiers in Psychology, 4*. Retrieved from doi:10.3389/fpsyg.2013.00433

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*, 1120-1135.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152-176.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117*, 1025-1054. doi:10.1037/a0020874

Yellott, J. I. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology, 15*, 109-144.

Yntema, D. B., & Mueser, G. E. (1962). Keeping track of variables that have few or many states. *Journal of Experimental Psychology, 63*, 391-395.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*, 233-236.

Table 1: Summary of Best-Fitting Models for All Experiments

Experiment	Filtering	Removal	Encoding	Deletion	C	A	F	R	E	D
1 CS	--	--	--	--	10.3	0.88	--	--	--	--
2 CS	a + c	c	c	--	13.4	0.88	0.58	16.7	0.59	--
3 CS	a + c	c	a + c	--	14.1	0.85	0.44	25.6	1.17	--
4 CS	a + c	c	a + c	--	15.0	0.81	0.61	16.3	2.6	--
5 MU	--	c	a + c	c	1.98	0.11	--	0.54	4.27	0.58
6 CS Young	a + c	c	c	--	12.8	1.16	0.56	18.1	0.83	--
6 CS Old	a + c	c	c	--	5.2	1.54	0.56	19.6	0.75	--
6 MU Young	--	a	c	c	11.7	0.87	--	-0.05	1.63	0.08
6 MU Old	--	a	c	c	7.1	1.34	--	0.06	0.82	0.11

The first columns indicate which of the two strength parameters (a and c) are affected by the model with the best (lowest) WAIC value in each experiment; the remaining columns present the posteriors of the sample means of the parameter estimates. CS = complex-span; MU = memory-updating.

Table 2. Model Recovery of Complex-Span Model Versions

Filter	Removal	Encoding	P(Rec)	P(false Rec)	M(ΔWAIC)	Med(ΔWAIC)
a	a	a	0.15	0.01	1.86	1.62
a	a	c	0.1	0	4.4	3.22
a	a	a+c	0.1	0	2.09	1.59
a	c	a	0.6	0.06	2.18	1.52
a	c	c	0.65	0.07	0.66	0.66
a	c	a+c	0.4	0.06	0.8	0.57
a	a+c	a	0.45	0.03	3.61	0.99
a	a+c	c	0.15	0.02	2.77	2.41
a	a+c	a+c	0.25	0.02	3.48	1.78
c	a	a	0.7	0.05	0.13	0.13
c	a	c	0.65	0.05	0.71	0.76
c	a	a+c	0.5	0.04	1.36	1.54
c	c	a	0.6	0.02	4.03	1.9
c	c	c	0.5	0.02	2.03	1
c	c	a+c	0.55	0.02	2.89	2.84
c	a+c	a	0.25	0.03	3.53	2.06
c	a+c	c	0.55	0.04	2.16	2.28
c	a+c	a+c	0.2	0.02	3.28	2.43
a+c	a	a	0.6	0.04	0.27	0.27
a+c	a	c	0.6	0.04	2.08	0.77
a+c	a	a+c	0.65	0.05	1.43	1.2
a+c	c	a	0.85	0.05	0.76	0.38
a+c	**c**	**c**	**0.7**	**0.05**	**3**	**2.23**
a+c	c	a+c	0.55	0.05	0.55	0.25
a+c	a+c	a	0.1	0.03	2.54	2.82
a+c	a+c	c	0.2	0.04	3.14	2.16
a+c	a+c	a+c	0.35	0.03	3.9	3.29

Note: The first three columns indicate the model version (which of the strength parameter each process parameter modifies); P(Rec) is the proportion of simulation runs recovering the true model (recovery hit rate); P(false Rec) is the proportion of simulation runs in which the model in a given row was falsely recovered when another model was true (recovery false-alarm rate); the last two columns give the mean (M) and median (Med) of the difference in WAIC between the true model and the best-fitting model, for those simulation runs in which the true model did not win the competition. The row printed in bold is the model version that fit best in Experiments 2 and 6.

Table 3. Model Recovery of Memory-Updating Model Versions

Time	Removal	Encoding	Deletion	P(Rec)	P(false Rec)	M(ΔWAIC)	Med(ΔWAIC)
1	a	a	a	1	0		
1	a	a	c	1	0		
1	a	a	a+c	1	0		
1	a	c	a	0.14	0	2.41	1.68
1	a	c	c	0.81	0	2.7	1.77
1	a	c	a+c	0.48	0	3.69	2.53
1	a	a+c	a	0.81	0	0.67	0.64
1	a	a+c	c	1	0		
1	a	a+c	a+c	0.9	0	4.02	4.02
1	c	a	a	1	0		
1	c	a	c	1	0		
1	c	a	a+c	1	0		
1	c	c	a	0.95	0	4.95	4.95
1	c	c	c	1	0		
1	c	c	a+c	1	0		
1	c	a+c	a	1	0		
1	c	a+c	c	1	0		
1	c	a+c	a+c	1	0		
1	a+c	a	a	1	0		
1	a+c	a	c	1	0		
1	a+c	a	a+c	1	0		
1	a+c	c	a	0.81	0	2.2	1.85
1	a+c	c	c	1	0		
1	a+c	c	a+c	0.95	0	0.98	0.98
1	a+c	a+c	a	1	0		
1	a+c	a+c	c	1	0		
1	a+c	a+c	a+c	1	0		
4	a	a	a	0.81	0	1.46	1.07
4	a	a	c	1	0		
4	a	a	a+c	0.95	0	1.28	1.28
4	a	c	a	0.14	0	2.07	2.19
4	**a**	**c**	**c**	**0.33**	**0**	**2.34**	**1.33**
4	a	c	a+c	0.19	0	2.02	1.14
4	a	a+c	a	0.33	0	2.14	2.1
4	a	a+c	c	0.95	0.01	4.63	4.63
4	a	a+c	a+c	0.67	0	1.6	1.84
4	c	a	a	1	0		
4	c	a	c	0.9	0	1.23	1.23
4	c	a	a+c	0.95	0	0.14	0.14
4	c	c	a	0.81	0.01	3.13	3.02
4	c	c	c	1	0		
4	c	c	a+c	0.86	0	2.36	3.4
4	c	a+c	a	1	0		
4	**c**	**a+c**	**c**	1	0		
4	c	a+c	a+c	1	0		
4	a+c	a	a	1	0		
4	a+c	a	c	1	0		

4	a+c	a	a+c	0.89	0	1.33	1.33
4	a+c	c	a	0.42	0	1.75	0.79
4	a+c	c	c	0.63	0.01	2.85	2.77
4	a+c	c	a+c	0.79	0.01	1.26	0.85
4	a+c	a+c	a	0.95	0	0.45	0.45
4	a+c	a+c	c	1	0		
4	a+c	a+c	a+c	1	0		

Note: The first four columns indicate the model version (Time=1 for using only CWI for removal, and only WCI for encoding; Time=4 for using both intervals for both processes; the other three columns show which of the strength parameter, *a* and *c*, each process parameter modifies); P(Rec) is the proportion of simulation runs recovering the true model (recovery hit rate); P(false Rec) is the proportion of simulation runs in which the model in a given row was falsely recovered when another model was true (recovery false-alarm rate); the last two columns give the mean (M) and median (Med) of the difference in WAIC between the true model and the best-fitting model, for those simulation runs in which the true model did not win the competition. The rows printed in bold are the model versions that fit best in Experiments 4 and 6, respectively.

Table 4: Discrete-State and Continuous-Strength Modeling Frameworks for Memory Tests

Model Type	Old-New Recognition	Selection from N Candidates	Continuous Reproduction
Discrete State	High-Threshold Models	MPT	Mixture Model
Continuous Strength	2-dimensional SDT with Gaussian noise	M^3	IMM

Table A1: Mean Rate of Time-Outs, Decision Errors, and Response Times in Experiments 2, 3, and 4 (Standard Deviations in Parentheses).

Condition	Timeouts Items	Errors Items	RT Items	Timeouts Distractors	Errors Distractors	RT Distractors
Experiment 2						
Short free time	.05 (.04)	.17 (.09)	1.06 (0.12)	.07 (.06)	.20 (.11)	1.07 (0.14)
Long free time	.06 (.06)	.16 (.10)	1.04 (0.12)	.09 (.08)	.19 (.10)	1.10 (0.13)
Experiment 3						
Pre-cued, short	.06 (.07)	.13 (.06)	1.03 (0.11)	.06 (.05)	.13 (.05)	1.08 (0.13)
Pre-cued, long	.07 (.07)	.15 (.12)	1.03 (0.15)	.08 (.07)	.12 (.08)	1.06 (0.16)
Post-cued, short	.09 (.09)	.15 (.09)	1.07 (0.17)	.07 (.09)	.13 (.10)	1.05 (0.17)
Post-cued, long	.09 (.08)	.11 (.07)	1.07 (0.15)	.07 (.06)	.13 (.10)	1.03 (0.13)
Experiment 4						
Pre-cued, short	.05 (.07)	.16 (.10)	1.02 (0.14)	.05 (.06)	.17 (.14)	1.05 (0.14)
Pre-cued, long	.03 (.03)	.15 (.09)	0.98 (0.11)	.07 (.06)	.13 (.11)	1.07 (0.15)
Post-cued, short	.07 (.08)	.17 (.11)	1.04 (0.16)	.05 (.06)	.15 (.13)	1.02 (0.15)
Post-cued, long	.04 (.05)	.12 (.11)	0.98 (0.13)	.04 (.05)	.12 (.10)	0.97 (0.13)
Experiment 6						
Short, young	.04 (.05)	.08 (0.4)	1.18 (0.20)	.04 (.05)	.10 (.05)	1.17 (0.17)
Long, young	.03 (.05)	.08 (.05)	1.15 (0.21)	.04 (.06)	.10 (.05)	1.17 (0.18)
Short, old	.04 (.04)	.09 (.13)	1.28 (0.16)	.05 (.05)	.10 (.13)	1.33 (0.17)
Long, old	.03 (.05)	.08 (.13)	1.24 (0.16)	.06 (.06)	.09 (.12)	1.35 (0.17)

Table A2: Numbers of Responses per Category in each Condition, Averaged over Participants, Experiments 2 and 6 (Standard Deviation in Parentheses)

Condition	Correct	Other Item	Distractor in Position	Other Distractor	NPL
	Experiment 2				
Short free time	59.89 (22.03)	19.74 (10.20)	8.81 (5.14)	9.67 (6.54)	3.19 (3.26)
Long free time	66.85 (20.28)	16.63 (10.21)	6.26 (3.69)	6.07 (6.70)	2.89 (4.12)
	Experiment 6				
Short free time, young	58.93 (21.98)	20.71 (11.57)	8.50 (5.17)	9.31 (6.53)	2.53 (3.41)
Long free time, young	69.26 (20.49)	16.41 (11.90)	6.17 (4.40)	6.23 (5.96)	1.94 (3.13)
Short free time, old	36.22 (18.63)	33.71 (11.37)	10.02 (3.91)	16.01 (8.21)	3.98 (3.23)
Long free time, old	45.96 (22.33)	29.82 (12.71)	8.90 (4.67)	12.67 (8.93)	2.65 (2.70)

Table A3: Numbers of Responses per Category in each Condition, Averaged over Participants, Experiments 3 and 4 (Standard Deviation in Parentheses)

Condition	Correct	Other Item	Distractor in Position	Other Distractor	NPL
	Experiment 3				
Pre-cued, short	33.79 (10.28)	10.79 (6.90)	1.42 (1.10)	3.63 (3.16)	1.38 (3.49)
Pre-cued, long	33.92 (10.46)	10.79 (7.37)	1.13 (1.51)	3.21 (3.23)	0.96 (1.68)
Post-cued, short	29.38 (12.98)	10.38 (6.57)	2.46 (2.70)	5.92 (4.90)	1.88 (2.23)
Post-cued, long	33.38 (9.94)	9.42 (7.14)	1.79 (1.35)	4.67 (3.71)	0.75 (0.99)
	Experiment 4				
Pre-cued, short	32.88 (11.11)	10.63 (7.95)	1.42 (0.97)	3.92 (3.91)	1.17 (1.49)
Pre-cued, long	36.96 (10.89)	8.92 (8.08)	0.75 (0.99)	2.54 (3.26)	0.83 (1.49)
Post-cued, short	28.50 (12.21)	11.58 (8.31)	2.67 (2.16)	6.13 (5.04)	1.13 (1.87)
Post-cued, long	36.88 (11.60)	8.13 (8.17)	1.71 (1.92)	2.67 (3.29)	0.63 (1.01)

Table A4: Numbers of Responses per Category in each Condition in the Memory Updating Task, Averaged over Participants, Experiments 5 and 6 (Standard Deviation in Parentheses)

Condition	Correct	Other Item	Distractor in Position	Other Distractor	NPL
	Experiment 5				
Short CWI, short WCI	15.55 (5.87)	3.25 (1.65)	6.05 (3.69)	5.50 (1.96)	1.65 (1.35)
Short CWI, long WCI	18.95 (5.31)	3.30 (1.75)	5.10 (2.99)	4.00 (2.36)	0.65 (0.81)
Long CWI, short WCI	18.40 (4.52)	3.40 (1.88)	5.60 (2.74)	3.55 (2.31)	1.05 (1.05)
Long CWI, long WCI	20.85 (5.64)	3.65 (2.03)	3.20 (2.48)	3.75 (2.61)	0.55 (0.83)
	Experiment 6				
Short CWI, short WCI, young	34.52 (9.39)	5.53 (6.07)	4.76 (2.79)	6.46 (3.34)	0.97 (1.31)
Short CWI, long WCI, young	38.53 (8.97)	4.74 (6.64)	4.37 (2.75)	3.99 (2.68)	0.62 (0.93)
Long CWI, short WCI, young	38.04 (9.18)	4.72 (6.91)	3.96 (2.67)	4.63 (3.04)	0.88 (1.35)
Long CWI, long WCI, young	39.97 (10.08)	3.65 (6.16)	4.09 (3.18)	3.62 (3.43)	0.91 (1.86)
Short CWI, short WCI, old	22.46 (11.78)	12.70 (10.20)	5.34 (3.05)	9.63 (3.93)	1.41 (1.23)

Short CWI, long WCI, old	26.61 (12.66)	11.02 (9.62)	5.36 (3.28)	7.78 (3.99)	1.24 (1.63)
Long CWI, short WCI, old	26.42 (12.82)	10.98 (10.43)	5.22 (2.69)	8.20 (4.54)	1.17 (2.07)
Long CWI, long WCI, old	26.68 (13.67)	11.52 (10.40)	5.20 (2.88)	7.44 (4.11)	1.15 (2.12)

Note: CWI= Cue-word interval; WCI = word-cue interval

Table A5. *Sample characteristics, Experiment 6.*

Measure	Young adults		Older adults	Comparisons between young adults (subsample) and older adults
	All	Subsample [a]		
Sample Size	68	24	59	-
Age (years)	24.4 (3.6)	23.9 (2.8)	70.8 (2.8)	
Gender (female/male)	57/11	18/6	25/34	
Mini Mental State (MMS)	-	29.50 (0.83)	29.03 (0.95)	$t(81) = 2.10$, $p < .05$
CERAD total score [b]	-	96.29 (8.58)	89.54 (8.47)	$t(81) = 3.28$, $p < .01$
Health				
physical index (standardized score) [c]	-	56.61 (5.73)	52.54 (5.69)	$t(81) = 2.95$, $p < .01$
mental index (standardized score) [c]	-	45.59 (7.91)	57.00 (4.37)	$t(81) = -8.40$, $p < .001$

Note. Standard deviations (SD) are given in parentheses.

[a] These participants already participated in a study from our lab (Rey-Mermet et al., 2016) and thus had background measures.

[b] This total score was computed as the sum score of the Boston Naming, figure drawing, word list learning, word list recall, word list recognition discriminability and verbal fluency (see Chandler et al., 2005, for the exact computation procedure).

[c] Higher scores indicate better health status.

Figure Captions

Figure 1: Top: Simulated data from basic M^3 of a serial recall task (list length 5), with an experimental manipulation (for instance, short vs. long presentation time) that selectively increases *a* (left), increases *c* (middle), or increases both *a* and *c* (right). We simulated 200 responses per condition from 50 subjects. Increasing *a* leads to more frequent recall of list items other than the correct one; increasing *c* leads to more correct recalls, and increasing both *a* and *c* results in increased recall of correct and other list items at the expense of not-presented lures (NPL). Bottom: Means of posteriors of individual subject parameters of change; Δ*a* represents the change in *a* between experimental conditions, and Δ*c* represents the change in *c*.

Figure 2: Top panel: Bars depict the observed mean proportion of responses in each category for the three conditions in Experiment 1. Model predictions (means and 95% confidence intervals of the mean posterior predictives for each participant) are presented as red dots for the unconstrained M^3 and blue dots for the constrained M^3. Bottom panel: Posterior probability distribution of the means of individual participant's parameter estimates of the unconstrained M^3. The black bar underneath each distribution marks the 95% highest-density interval (HDI; Kruschke, 2011). The Δa_1 and Δc_1 parameters for the *old-reordered* condition have black bars to indicate their HDIs, and the Δa_2 and Δc_2 parameters for the *old-same* condition have red bars.

Figure 3: Posterior distributions of means of individual participants' parameter values from the constrained M^3 (i.e., Δ*c* applied only to the *old-same* condition, and Δa equal for *old-reordered*, and *old-same* conditions) applied to Experiment 1. The black bar underneath each distribution marks the 95% HDI.

Figure 4: Goodness of fit of the 64 model versions applied to Experiment 2. Gray scale represents difference of each model's WAIC from the smallest (best) WAIC. Darker shading

means larger ΔWAIC values; steps are logarithmically spaced. The smallest WAIC value was 1092.4.

Figure 5: Mean probability of choosing a word from each of the five response categories (bars), with predictions from the best-fitting M^3 model (red dots) for Experiment 2. P(choice) is the probability of choosing each individual word in a given category, so for response categories with more than one word, the proportion of responses in that category was divided by the number of words in the candidate set belonging to that category. Note the scale difference between the first and subsequent panels! Error bars are 95% confidence intervals for within-subjects comparisons (Bakeman & McArthur, 1996). Model predictions are derived from the means of the posterior predictives for each participant and category.

Figure 6: Posterior probability densities of the means of parameter estimates across participants for the best-fitting M^3 model of Experiment 2. The distributions were constructed by taking the average of all participants' parameter values at each MCMC sampling step, then plotting a smoothed histogram of these averages across all MCMC steps. Thick horizontal bars represent the 95% highest-density interval (HDI) (Kruschke, 2011).

Figure 7: Goodness of fit of the 64 model versions applied to Experiment 3. Gray scale represents difference of each model's WAIC from the smallest (best) WAIC. Darker shading means larger ΔWAIC values; steps are logarithmically spaced. The smallest WAIC value was 1737.9.

.

Figure 8: Mean probability of choosing a word from each of the five response categories (bars), with predictions from the best-fitting M^3 model (red dots) for Experiment 3. See legend of Figure 5 for details. Pre = pre-cued condition, Post = post-cued condition, Short = short free time, Long = long free time.

Figure 9. Posterior probability densities of the means of parameter estimates across participants for the best-fitting M^3 model of Experiment 3. See legend of Figure 6 for details.

Figure 10. Goodness of fit of the 64 model versions applied to Experiment 4. Gray scale represents difference of each model's WAIC from the smallest (best) WAIC. Darker shading means larger ΔWAIC values; steps are logarithmically spaced. The smallest WAIC value was 1651.1.

Figure 11: Mean probability of choosing a word from each of the five response categories (bars), with predictions from the best-fitting M^3 model (red dots) for Experiment 4. See legend of Figure 5 for details. Pre = pre-cued condition, Post = post-cued condition, Short = short free time, Long = long free time.

Figure 12. Posterior probability densities of the means of parameter estimates across participants for the best-fitting M^3 model of Experiment 4. See legend of Figure 6 for details.

Figure 13. Goodness of fit of the 256 model versions applied to Experiment 5. Gray scale represents difference of each model's WAIC from the smallest (best) WAIC. Darker shading means larger ΔWAIC values; steps are logarithmically spaced. The smallest WAIC value was 1249.4.

Figure 14: Mean probability of choosing a word from each of the five response categories (bars), with predictions from the best-fitting M^3 model (red dots) for Experiment 5. Note the scale difference between the first and subsequent panels! See legend of Figure 5 for details. CWI = cue-word interval, WCI = word-cue interval; short intervals are represented by "–" and long intervals by "+".

Figure 15. Posterior probability densities of the means of parameter estimates across participants for the best-fitting M^3 model of Experiment 5. See legend of Figure 6 for details.

Figure 16: Goodness of fit of the 64 model versions applied to the complex-span task of Experiment 6. Gray scale represents difference of each model's WAIC from the smallest (best) WAIC. Darker shading means larger ΔWAIC values; steps are logarithmically spaced. The smallest WAIC value was 4666.7.

Figure 17: Mean probability of choosing a word from each of the five response categories (bars), with predictions from the best-fitting M^3 model (red dots) for the complex-span task in Experiment 6. See legend of Figure 5 for details.

Figure 18. Posterior probability densities of the means of parameter estimates across participants for the best-fitting M^3 model for the complex-span task of Experiment 6. See legend of Figure 6 for details. Posteriors of young adults are marked by black HDI bars; those of old adults by red HDI bars.

Figure 19: Posteriors of the differences between young and old population-level parameters for complex span, with 95% HDIs (black horizontal bars). Green text gives the proportions of the posteriors that fall below and above zero. Note that the population-level parameter for *f* is on a logit-scale.

Figure 20: Goodness of fit of the 256 model versions applied to the updating task of Experiment 6. Gray scale represents difference of each model's WAIC from the smallest (best) WAIC. Darker shading means larger ΔWAIC values; steps are logarithmically spaced. The smallest WAIC value was 8686.3.

Figure 21: Mean probability of choosing a word from each of the five response categories (bars), with predictions from the best-fitting M^3 model (red dots) for the memory-updating task in Experiment 6. See legend of Figure 5 for details. CWI = cue-word interval, WCI = word-cue interval; short intervals are represented by "–" and long intervals by "+".

Figure 22. Posterior probability densities of the means of parameter estimates across participants for the best-fitting M^3 model for the memory-updating task of Experiment 6. See legend of Figure 6 for details. Posteriors of young adults are marked by black HDI bars; those of old adults by red HDI bars.

Figure 23: Posteriors of the differences between young and old population-level parameters for memory updating, with 95% HDIs (black horizontal bars). Green text gives the proportions of the posteriors that fall below and above zero. Note that the population-level parameter for *d* is on a logit-scale.

Figure 24: Parameter recovery simulation of the complex-span model, simulating data from Experiment 2, varying the group mean of parameter *c*. True values of *c* are plotted on the x-axis, the estimated values (means and 95% HDIs) for the five model parameters are plotted along the y-axis. The dotted lines represent the true parameter value for the parameter plotted in each panel.

Figure 25: Parameter recovery simulation of the complex-span model, simulating data from Experiment 2, varying the group mean of parameter *a*. True values of *a* are plotted on the x-axis, the estimated values (means and 95% HDIs) for the five model parameters are plotted along the y-axis. The dotted lines represent the true parameter value for the parameter plotted in each panel.

Figure 26: Parameter recovery simulation of the complex-span model, simulating data from Experiment 2, varying the group mean of parameter *f*. True values of *f* are plotted on the x-axis, the estimated values (means and 95% HDIs) for the five model parameters are plotted along the y-axis. The dotted lines represent the true parameter value for the parameter plotted in each panel.

Figure 27: Parameter recovery simulation of the complex-span model, simulating data from Experiment 2, varying the group mean of parameter *r*. True values of *r* are plotted on the x-axis, the estimated values (means and 95% HDIs) for the five model parameters are plotted along the y-axis. The dotted lines represent the true parameter value for the parameter plotted in each panel.

Figure 28: Parameter recovery simulation of the complex-span model, simulating data from Experiment 2, varying the group mean of parameter *e*. True values of *e* are plotted on the x-axis, the estimated values (means and 95% HDIs) for the five model parameters are plotted along the y-axis. The dotted lines represent the true parameter value for the parameter plotted in each panel.

Figure 29: Parameter recovery simulation of the memory-updating model, simulating data from Experiment 5, varying the group mean of parameter *c*. True values of *c* are plotted on the x-axis, the estimated values (means and 95% HDIs) for the five model parameters are plotted along the y-axis. The dotted lines represent the true parameter value for the parameter plotted in each panel.

Figure 30: Parameter recovery simulation of the memory-updating model, simulating data from Experiment 5, varying the group mean of parameter *a*. True values of *a* are plotted on the x-axis, the estimated values (means and 95% HDIs) for the five model parameters are plotted along the y-axis. The dotted lines represent the true parameter value for the parameter plotted in each panel.

Figure 31: Parameter recovery simulation of the memory-updating model, simulating data from Experiment 4, varying the group mean of parameter *d*. True values of *d* are plotted on the x-axis, the estimated values (means and 95% HDIs) for the five model parameters are

plotted along the y-axis. The dotted lines represent the true parameter value for the parameter plotted in each panel.

Figure 32: Parameter recovery simulation of the memory-updating model, simulating data from Experiment 5, varying the group mean of parameter *r*. True values of *r* are plotted on the x-axis, the estimated values (means and 95% HDIs) for the five model parameters are plotted along the y-axis. The dotted lines represent the true parameter value for the parameter plotted in each panel.

Figure 33: Parameter recovery simulation of the memory-updating model, simulating data from Experiment 5, varying the group mean of parameter *e*. True values of *e* are plotted on the x-axis, the estimated values (means and 95% HDIs) for the five model parameters are plotted along the y-axis. The dotted lines represent the true parameter value for the parameter plotted in each panel.

Figure A1: Predicted proportions of correct responses from the M^3 for the complex-span task in Experiments 2 and 6 (black = short free time; red = long free time), for variations of parameter values. Data come from five sets of simulations, each varying one parameter (*a*, *c*, *f*, *e*, or *r*) over eight values while holding the other parameters at an intermediate value (which was obtained by rounding the mean posterior estimate from Experiment 2). Values for a and c were divided by 5 for Luce(exp(A)) and by 7 for the SDT model to bring predicted accuracy into the same range as the other models. For LBA and LCA we set parameters of the accumulator models to values reported in the literature, adjusting them where necessary to bring accuracy into the same range as the other decision rules. Further simulations with the LBA showed that the predicted pattern does not qualitatively change with other parameter values for the accumulator model.

Figure A2: Probability of choosing each individual response candidate as a function of their position in the list relative to the position of the correct item; negative relative positions are positions earlier in the list than that of the correct item, and positive relative positions come later. Panels on the left

show choices of other list items (a. k. a. transpositions), panels on the right show distractor intrusions. The red lines represent the probability of choosing each not-presented lure (NPL).

Figure 1

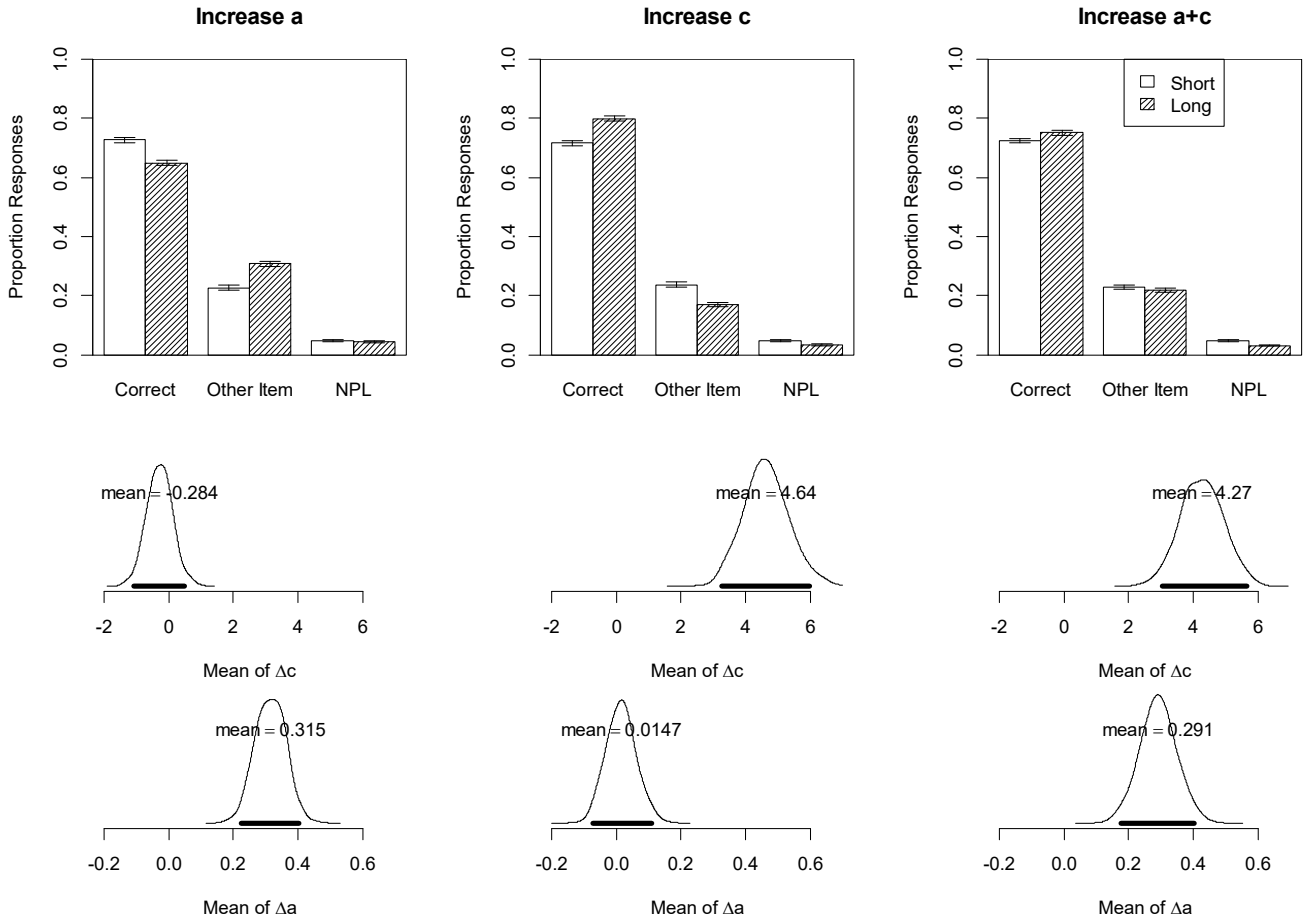
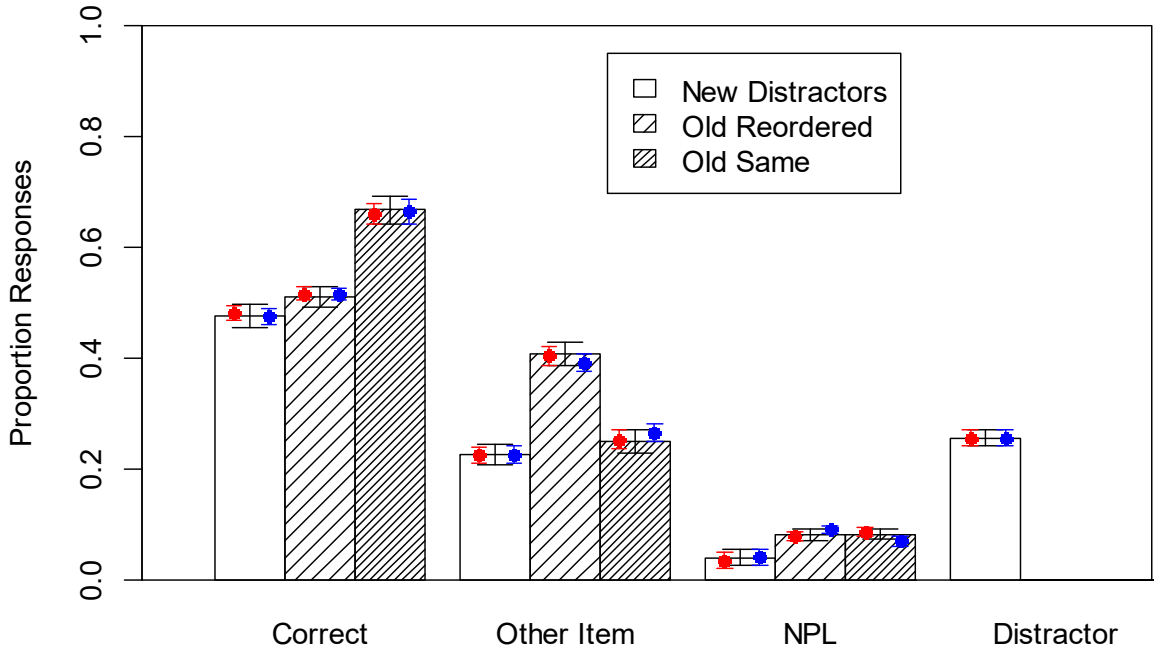

Increase a
Increase c
Increase a+c
Proportion Responses
0.0
0.2
0.4
0.6
0.8
1.0
Correct
Other Item
NPL
Short
Long
mean = -0.284
mean = 4.64
mean = 4.27
-2
0
2
4
6
Mean of Δc
mean = 0.315
mean = 0.0147
mean = 0.291
-0.2
0.0
0.2
0.4
0.6
Mean of Δa

Figure 2


Proportion Responses
0.0
0.2
0.4
0.6
0.8
1.0
New Distractors
Old Reordered
Old Same
Correct
Other Item
NPL
Distractor

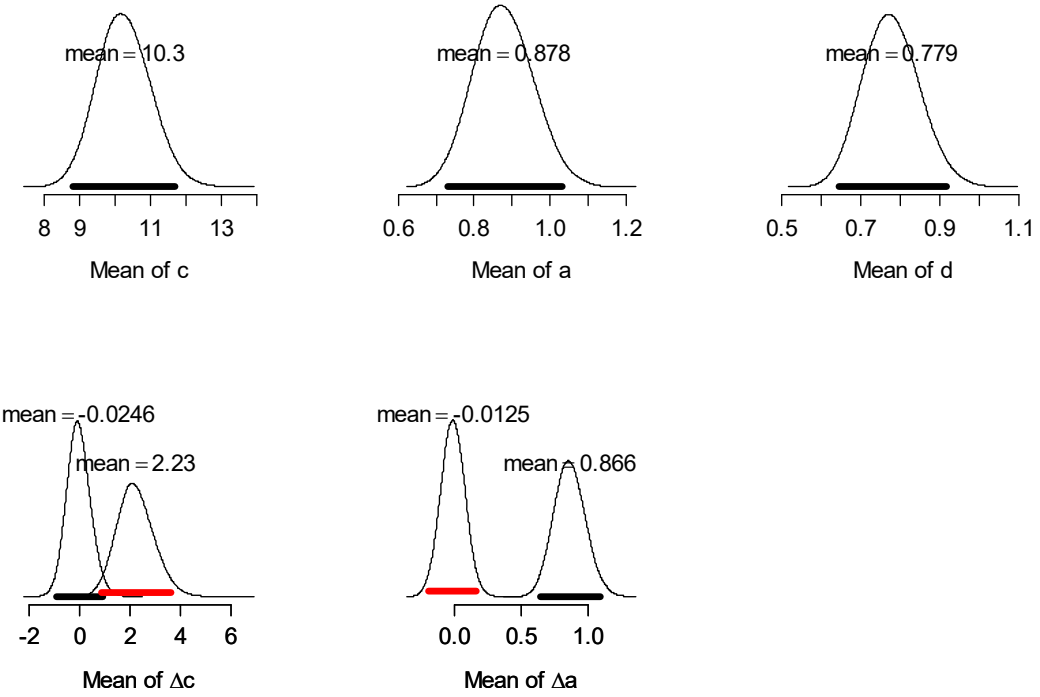

mean = 10.3
8 9 11 13
Mean of c
mean = 0.878
0.6 0.8 1.0 1.2
Mean of a
mean = 0.779
0.5 0.7 0.9 1.1
Mean of d
mean = -0.0246
mean = 2.23
-2 0 2 4 6
Mean of Δc
mean = -0.0125
mean = 0.866
0.0 0.5 1.0
Mean of Δa

Figure 3

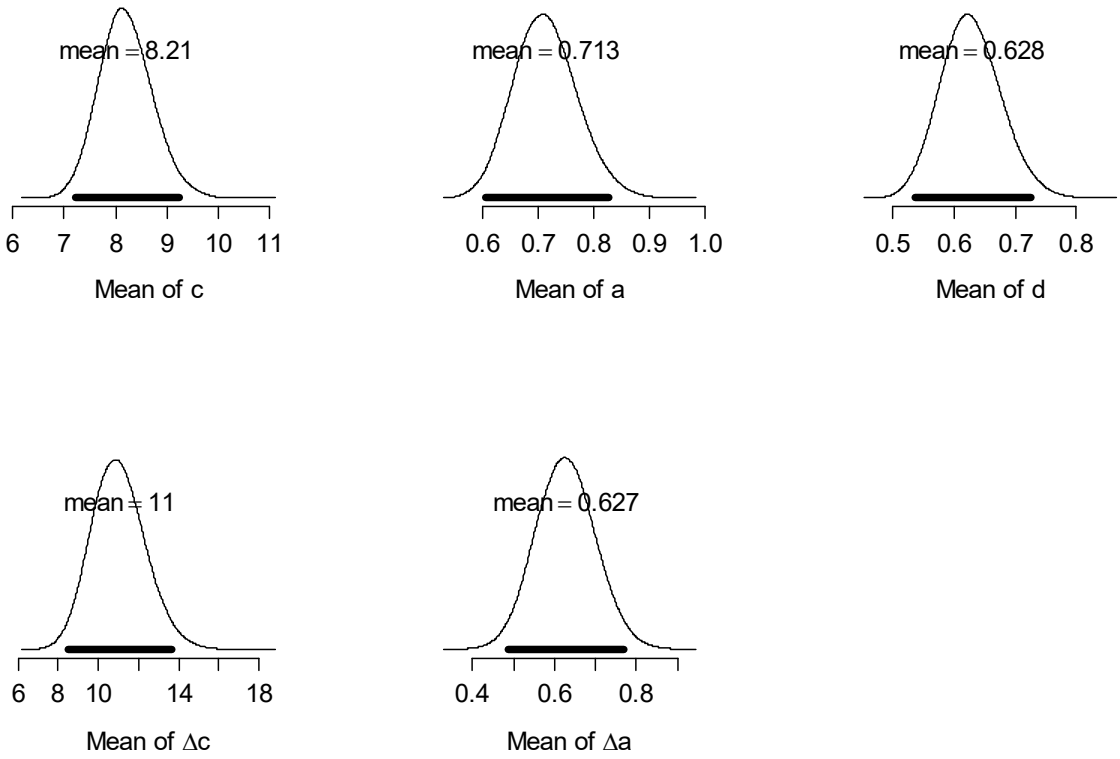

mean = 8.21
6 7 8 9 10 11
Mean of c
mean = 0.713
0.6 0.7 0.8 0.9 1.0
Mean of a
mean = 0.628
0.5 0.6 0.7 0.8
Mean of d
mean = 11
6 8 10 14 18
Mean of Δc
mean = 0.627
0.4 0.6 0.8
Mean of Δa

Figure 4.

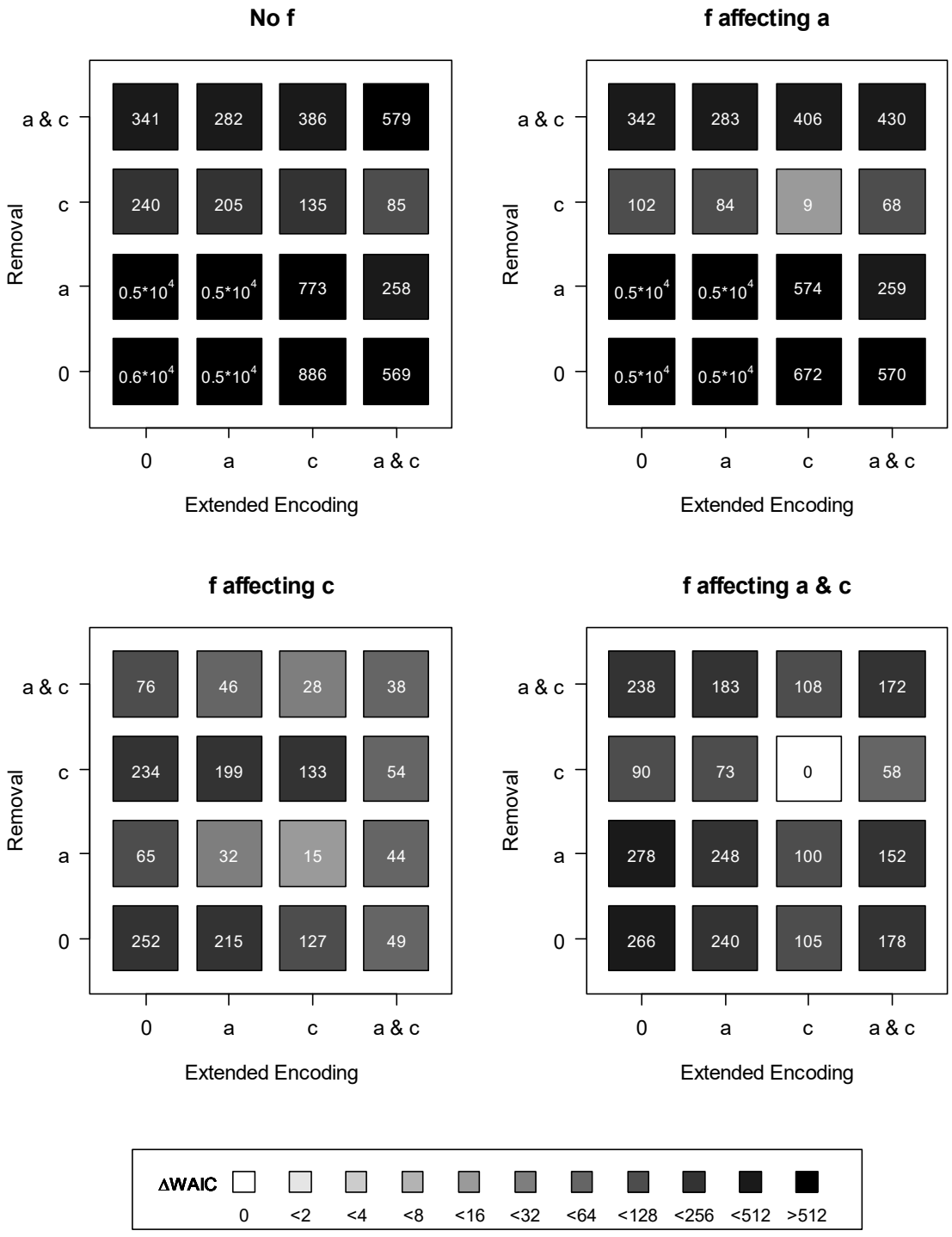

No f
341 282 386 579
240 205 135 85
0.5*10^4 0.5*10^4 773 258
0.6*10^4 0.5*10^4 886 569
f affecting a
342 283 406 430
102 84 9 68
0.5*10^4 0.5*10^4 574 259
0.5*10^4 0.5*10^4 672 570
f affecting c
76 46 28 38
234 199 133 54
65 32 15 44
252 215 127 49
f affecting a & c
238 183 108 172
90 73 0 58
278 248 100 152
266 240 105 178
Removal
a & c
c
a
0
0 a c a & c
Extended Encoding
ΔWAIC
0 <2 <4 <8 <16 <32 <64 <128 <256 <512 >512

Figure 5

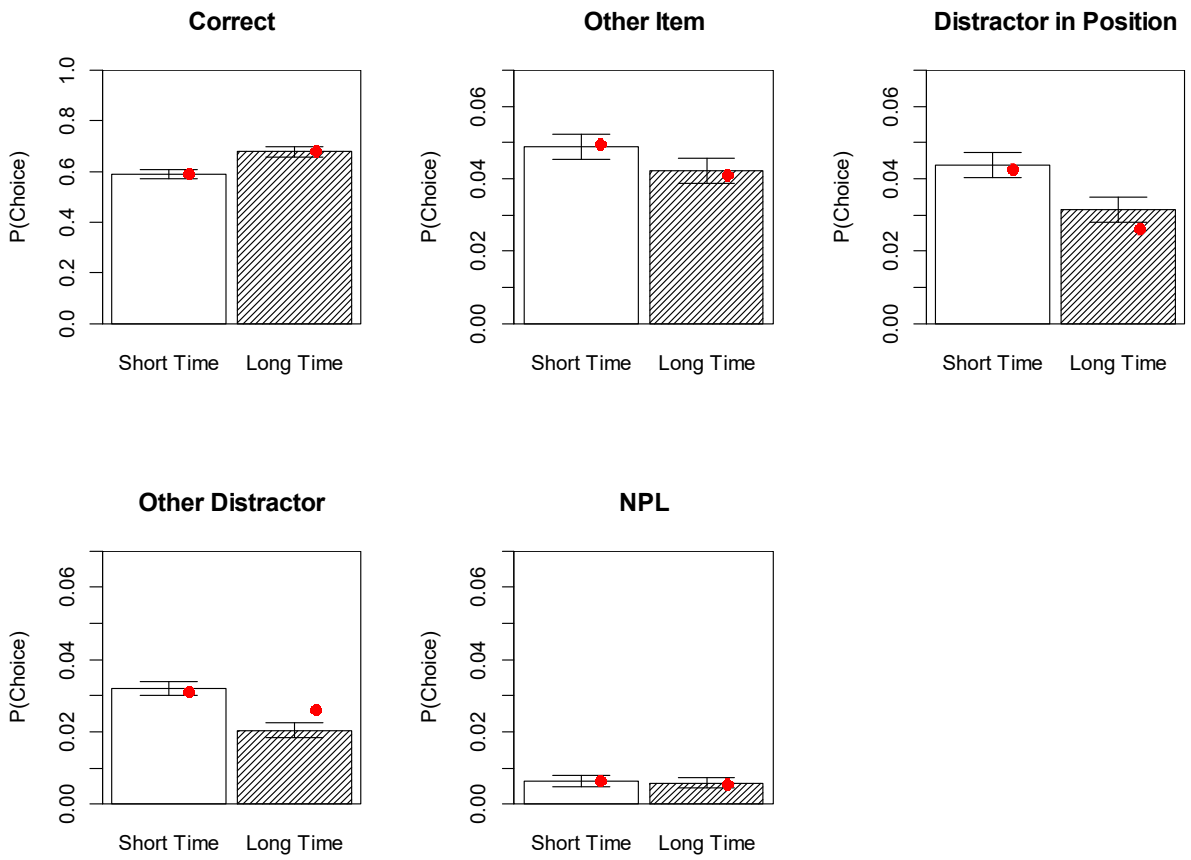
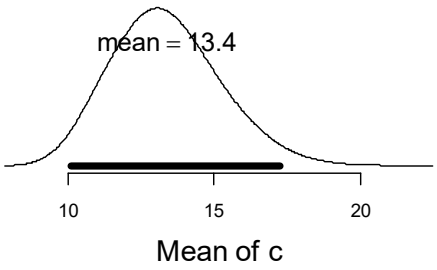
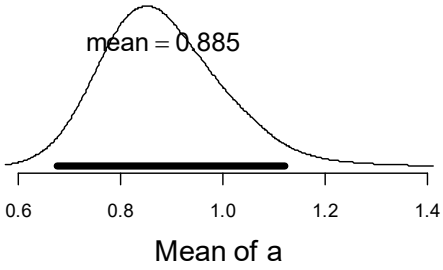
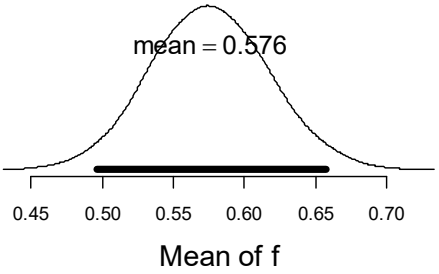
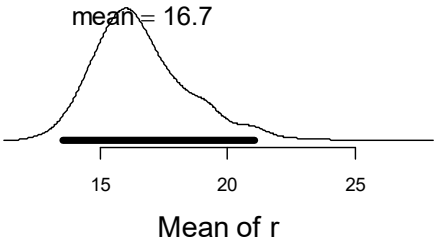

Correct
Other Item
Distractor in Position
Other Distractor
NPL
P(Choice)
Short Time
Long Time
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06

Figure 6


mean = 13.4
10
15
20
Mean of c


mean = 0.885
0.6
0.8
1.0
1.2
1.4
Mean of a


mean = 0.576
0.45
0.50
0.55
0.60
0.65
0.70
Mean of f


mean = 16.7
15
20
25
Mean of r

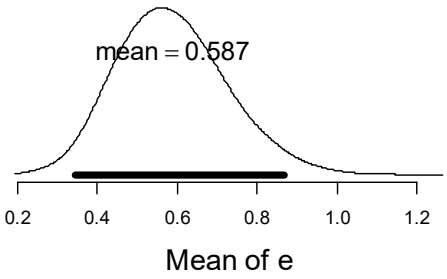

mean = 0.587
0.2
0.4
0.6
0.8
1.0
1.2
Mean of e

Figure 7

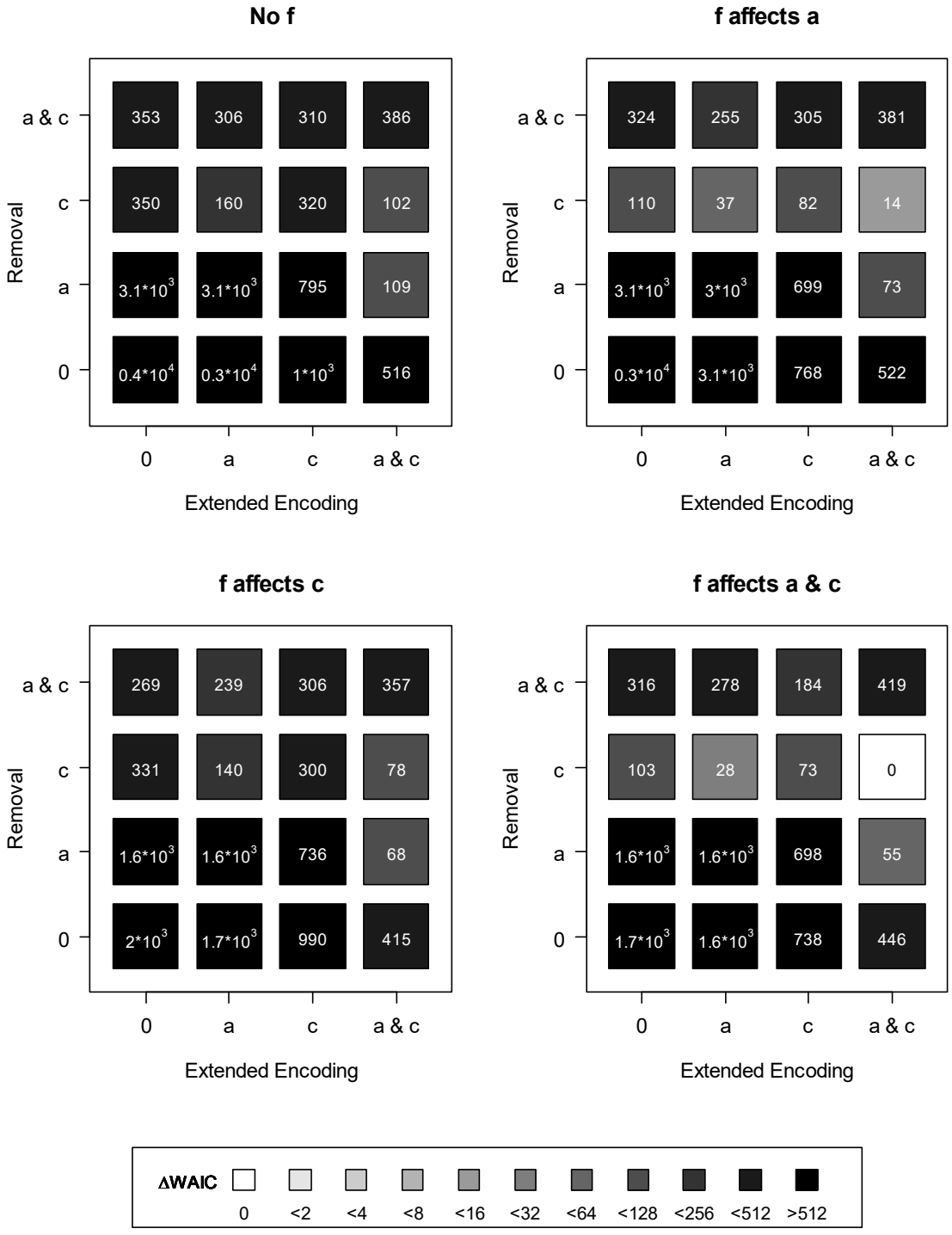

No f
353 306 310 386
350 160 320 102
3.1*10^3 3.1*10^3 795 109
0.4*10^4 0.3*10^4 1*10^3 516
f affects a
324 255 305 381
110 37 82 14
3.1*10^3 3*10^3 699 73
0.3*10^4 3.1*10^3 768 522
f affects c
269 239 306 357
331 140 300 78
1.6*10^3 1.6*10^3 736 68
2*10^3 1.7*10^3 990 415
f affects a & c
316 278 184 419
103 28 73 0
1.6*10^3 1.6*10^3 698 55
1.7*10^3 1.6*10^3 738 446
Removal
a & c
c
a
0
Extended Encoding
ΔWAIC
0 <2 <4 <8 <16 <32 <64 <128 <256 <512 >512

Figure 8

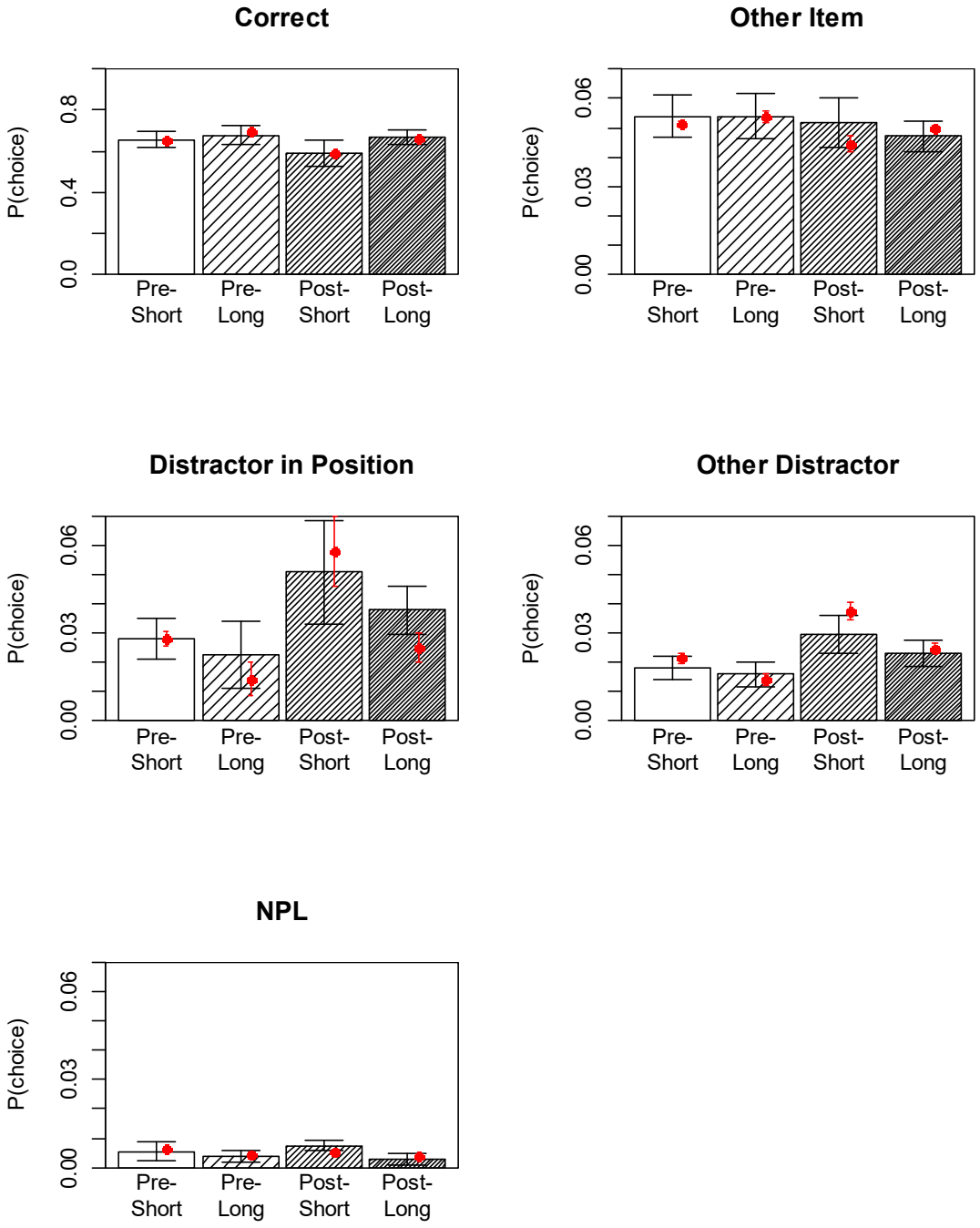
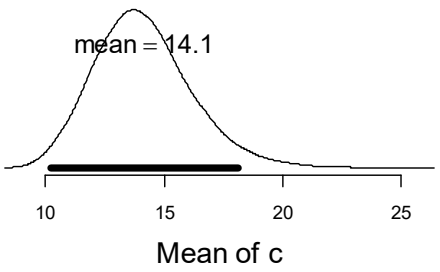
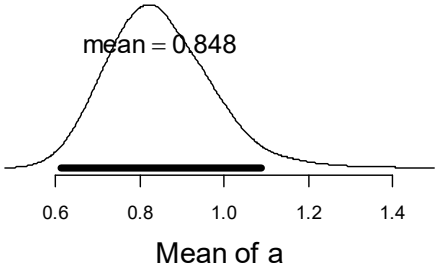
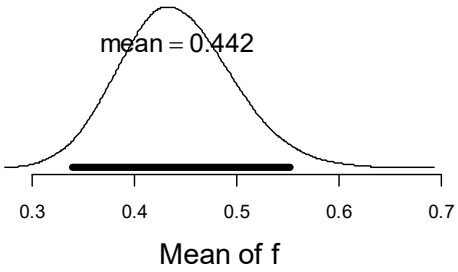
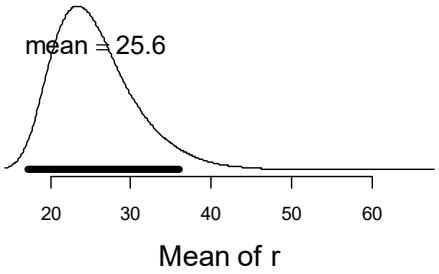

Correct
P(choice)
0.0
0.4
0.8
Pre-Short
Pre-Long
Post-Short
Post-Long
Other Item
P(choice)
0.00
0.03
0.06
Pre-Short
Pre-Long
Post-Short
Post-Long
Distractor in Position
P(choice)
0.00
0.03
0.06
Pre-Short
Pre-Long
Post-Short
Post-Long
Other Distractor
P(choice)
0.00
0.03
0.06
Pre-Short
Pre-Long
Post-Short
Post-Long
NPL
P(choice)
0.00
0.03
0.06
Pre-Short
Pre-Long
Post-Short
Post-Long

Figure 9


mean = 14.1
10
15
20
25
Mean of c


mean = 0.848
0.6
0.8
1.0
1.2
1.4
Mean of a


mean = 0.442
0.3
0.4
0.5
0.6
0.7
Mean of f


mean = 25.6
20
30
40
50
60
Mean of r

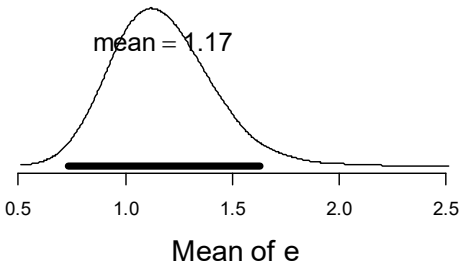

mean = 1.17
0.5
1.0
1.5
2.0
2.5
Mean of e

Figure 10

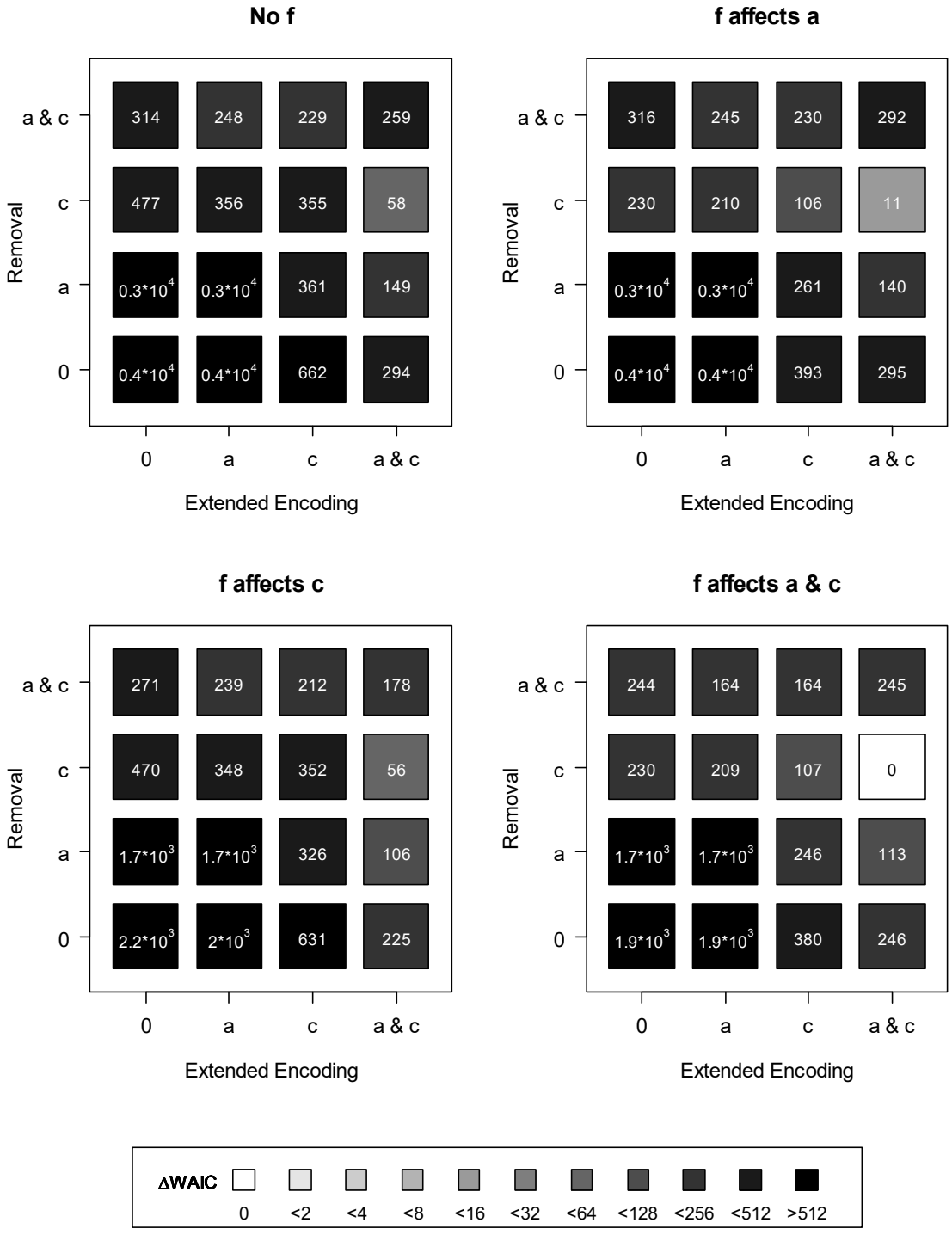

No f
f affects a
f affects c
f affects a & c
Removal
a & c
c
a
0
Extended Encoding
0
a
c
a & c
314 248 229 259
477 356 355 58
0.3*10^4 0.3*10^4 361 149
0.4*10^4 0.4*10^4 662 294
316 245 230 292
230 210 106 11
0.3*10^4 0.3*10^4 261 140
0.4*10^4 0.4*10^4 393 295
271 239 212 178
470 348 352 56
1.7*10^3 1.7*10^3 326 106
2.2*10^3 2*10^3 631 225
244 164 164 245
230 209 107 0
1.7*10^3 1.7*10^3 246 113
1.9*10^3 1.9*10^3 380 246
ΔWAIC
0 <2 <4 <8 <16 <32 <64 <128 <256 <512 >512

Figure 11

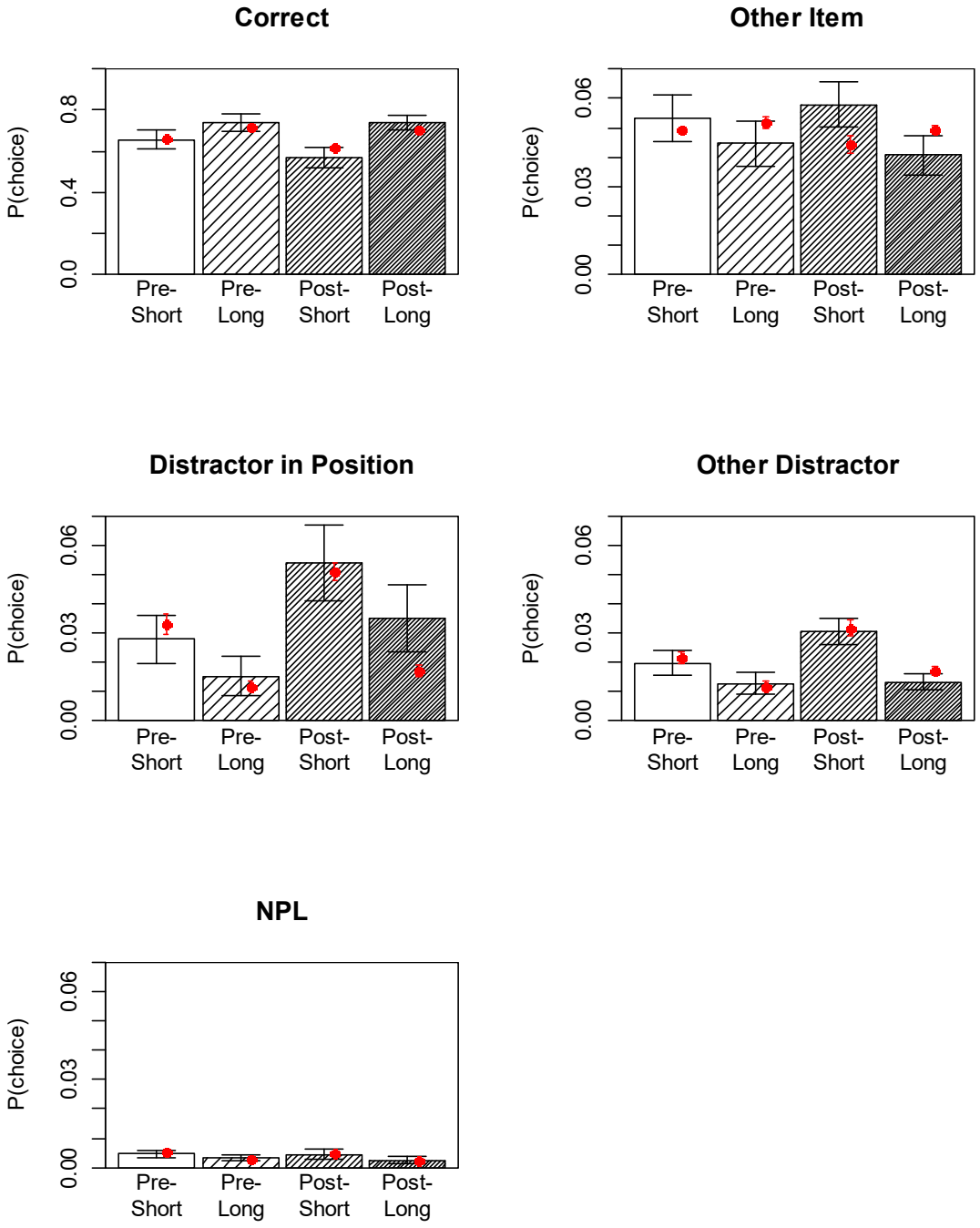

Correct
P(choice)
0.0
0.4
0.8
Pre-Short
Pre-Long
Post-Short
Post-Long
Other Item
P(choice)
0.00
0.03
0.06
Pre-Short
Pre-Long
Post-Short
Post-Long
Distractor in Position
P(choice)
0.00
0.03
0.06
Pre-Short
Pre-Long
Post-Short
Post-Long
Other Distractor
P(choice)
0.00
0.03
0.06
Pre-Short
Pre-Long
Post-Short
Post-Long
NPL
P(choice)
0.00
0.03
0.06
Pre-Short
Pre-Long
Post-Short
Post-Long

Figure 12

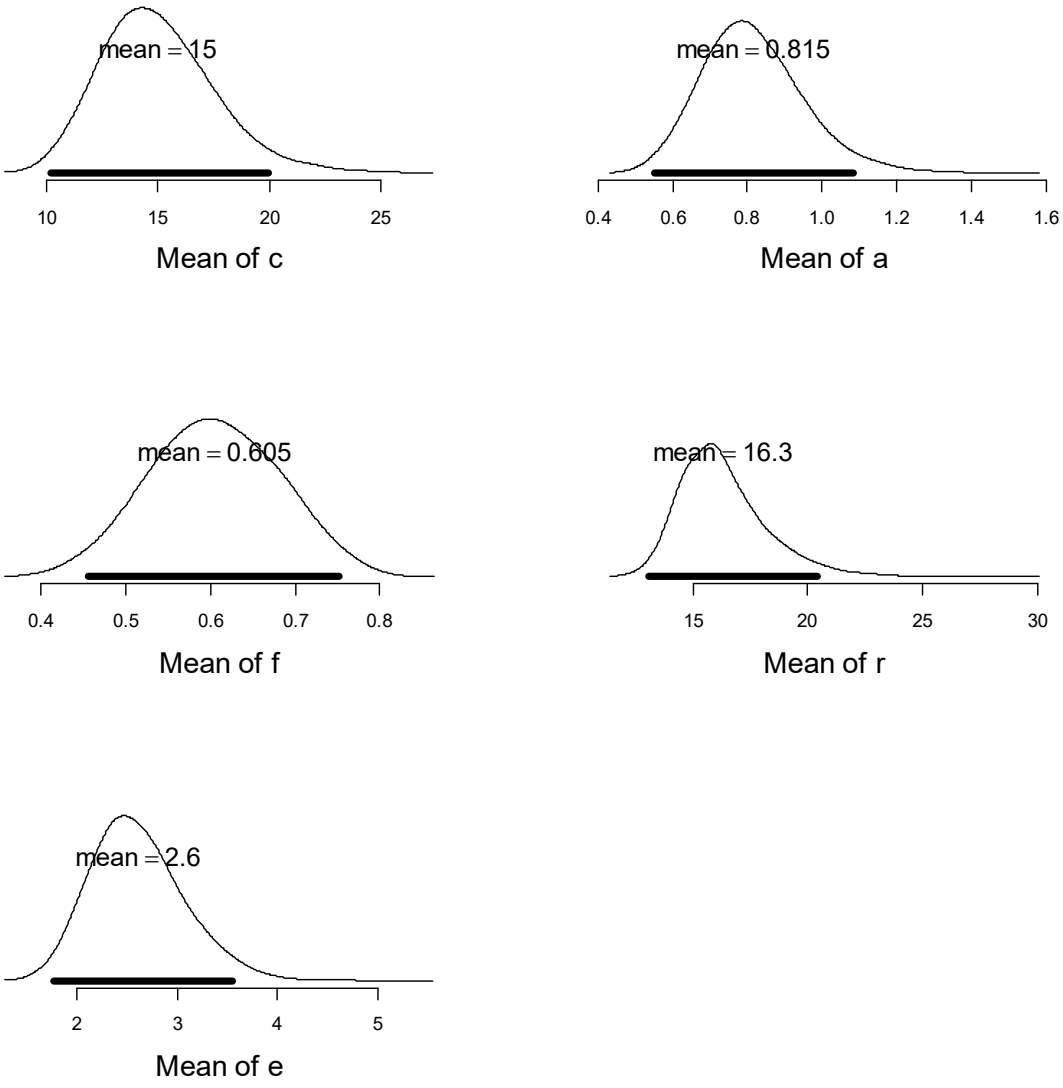

mean = 15
10
15
20
25
Mean of c
mean = 0.815
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Mean of a
mean = 0.605
0.4
0.5
0.6
0.7
0.8
Mean of f
mean = 16.3
15
20
25
30
Mean of r
mean = 2.6
2
3
4
5
Mean of e

Figure 13

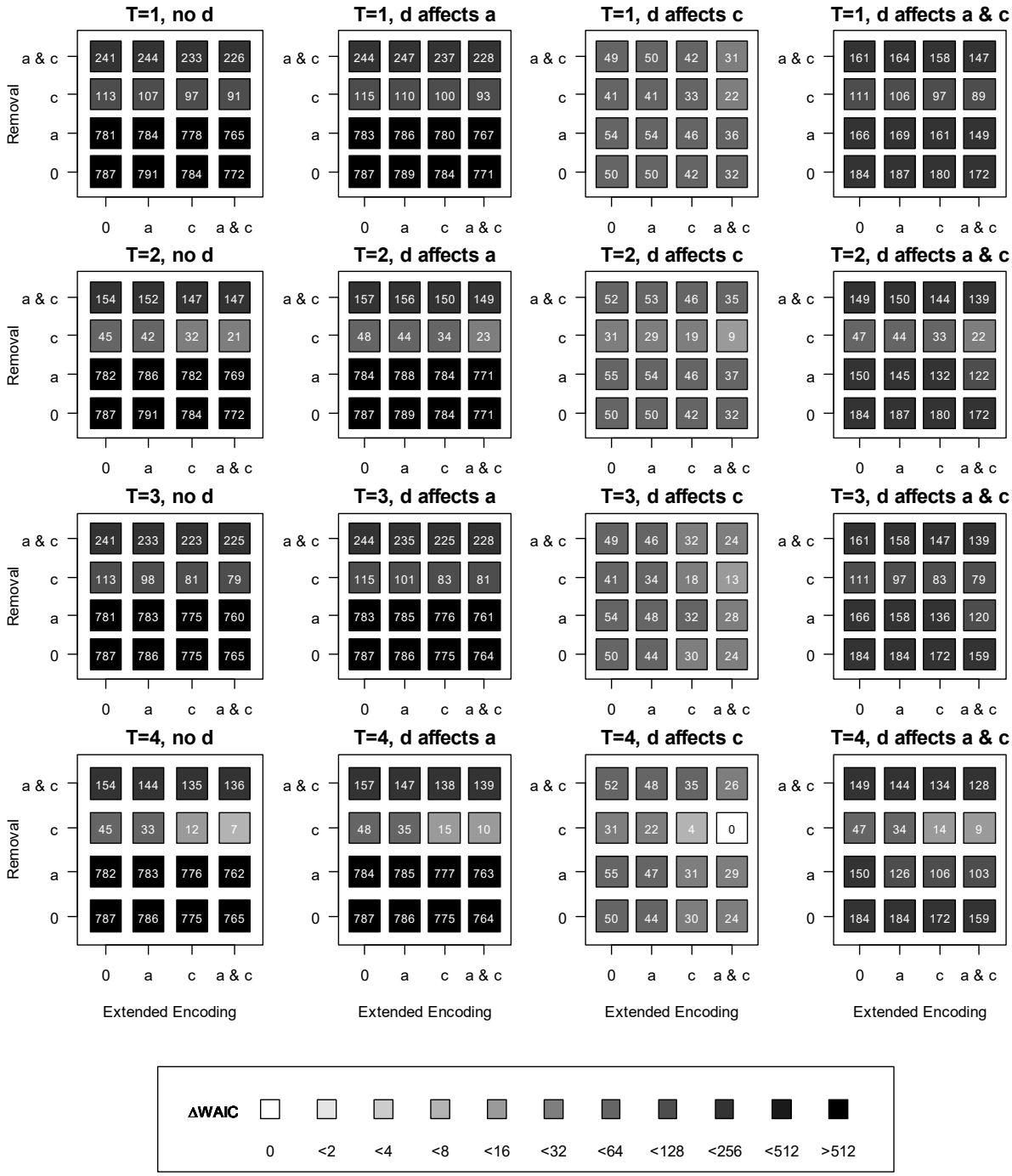

T=1, no d
241 244 233 226
113 107 97 91
781 784 778 765
787 791 784 772
T=1, d affects a
244 247 237 228
115 110 100 93
783 786 780 767
787 789 784 771
T=1, d affects c
49 50 42 31
41 41 33 22
54 54 46 36
50 50 42 32
T=1, d affects a & c
161 164 158 147
111 106 97 89
166 169 161 149
184 187 180 172
T=2, no d
154 152 147 147
45 42 32 21
782 786 782 769
787 791 784 772
T=2, d affects a
157 156 150 149
48 44 34 23
784 788 784 771
787 789 784 771
T=2, d affects c
52 53 46 35
31 29 19 9
55 54 46 37
50 50 42 32
T=2, d affects a & c
149 150 144 139
47 44 33 22
150 145 132 122
184 187 180 172
T=3, no d
241 233 223 225
113 98 81 79
781 783 775 760
787 786 775 765
T=3, d affects a
244 235 225 228
115 101 83 81
783 785 776 761
787 786 775 764
T=3, d affects c
49 46 32 24
41 34 18 13
54 48 32 28
50 44 30 24
T=3, d affects a & c
161 158 147 139
111 97 83 79
166 158 136 120
184 184 172 159
T=4, no d
154 144 135 136
45 33 12 7
782 783 776 762
787 786 775 765
T=4, d affects a
157 147 138 139
48 35 15 10
784 785 777 763
787 786 775 764
T=4, d affects c
52 48 35 26
31 22 4 0
55 47 31 29
50 44 30 24
T=4, d affects a & c
149 144 134 128
47 34 14 9
150 126 106 103
184 184 172 159
Removal
a & c
c
a
0
0 a c a & c
Extended Encoding
ΔWAIC
0 <2 <4 <8 <16 <32 <64 <128 <256 <512 >512

Figure 14

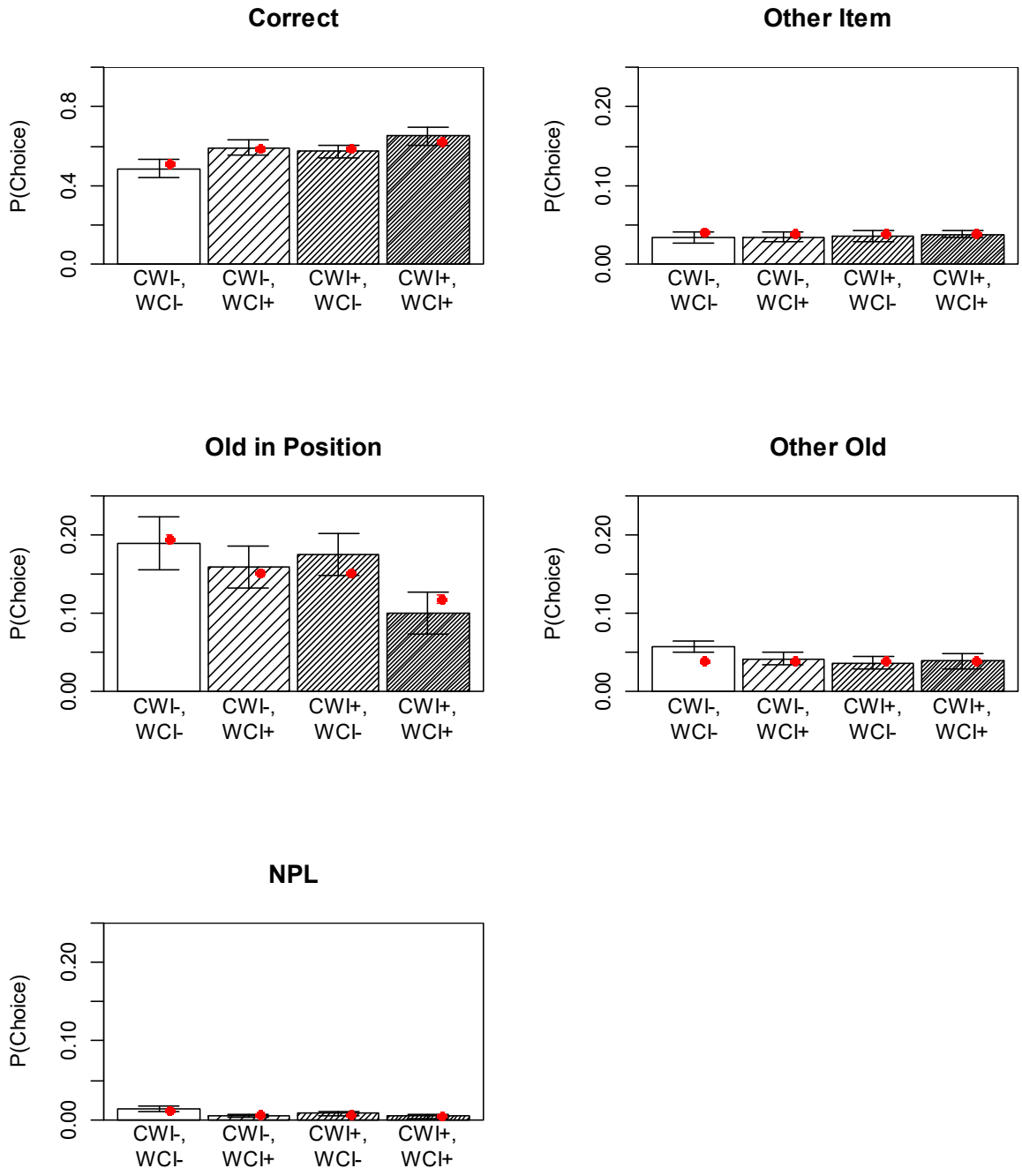

Correct
P(Choice)
0.0
0.4
0.8
CWI-, WCI-
CWI-, WCI+
CWI+, WCI-
CWI+, WCI+
Other Item
P(Choice)
0.00
0.10
0.20
CWI-, WCI-
CWI-, WCI+
CWI+, WCI-
CWI+, WCI+
Old in Position
P(Choice)
0.00
0.10
0.20
CWI-, WCI-
CWI-, WCI+
CWI+, WCI-
CWI+, WCI+
Other Old
P(Choice)
0.00
0.10
0.20
CWI-, WCI-
CWI-, WCI+
CWI+, WCI-
CWI+, WCI+
NPL
P(Choice)
0.00
0.10
0.20
CWI-, WCI-
CWI-, WCI+
CWI+, WCI-
CWI+, WCI+

Figure 15

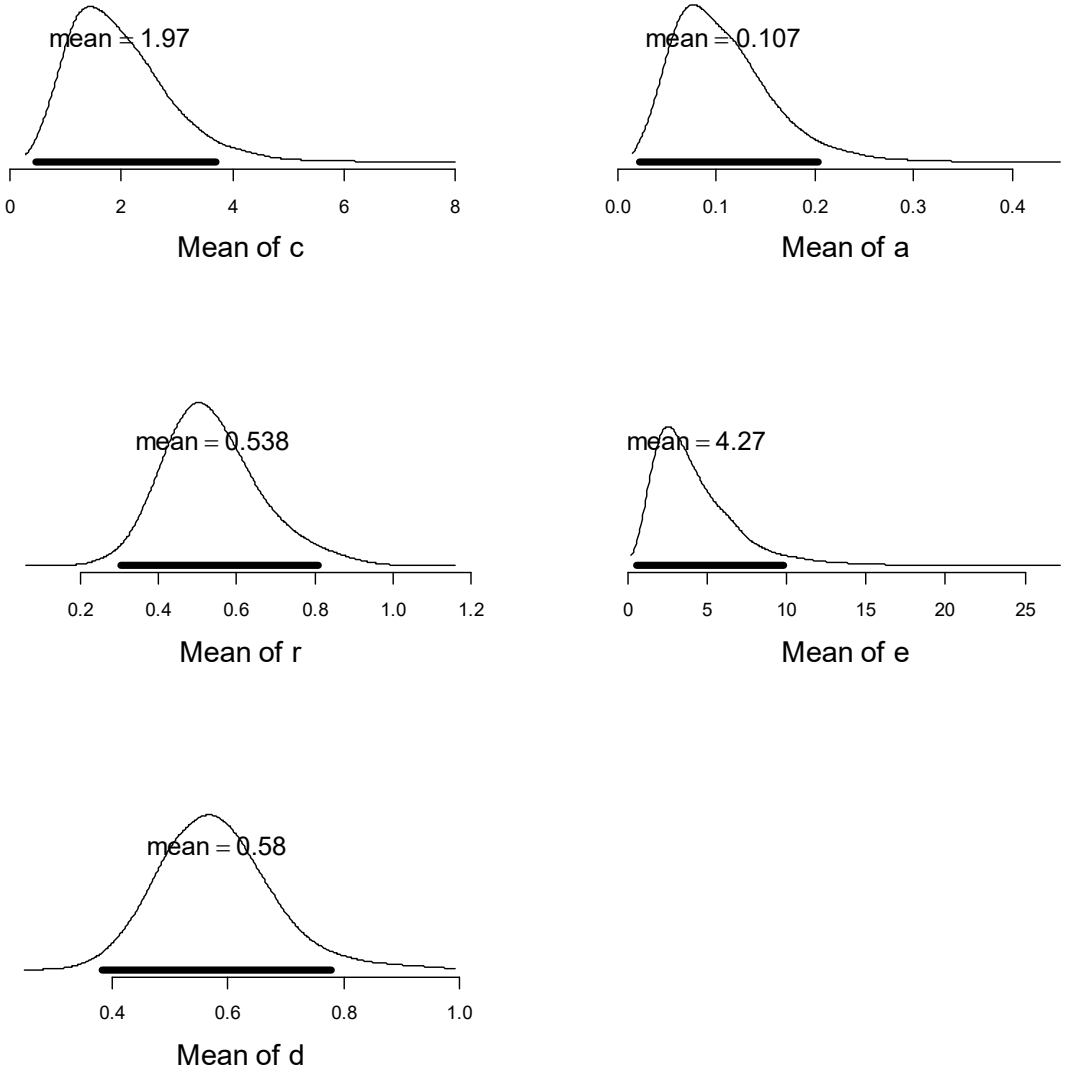

mean = 1.97
0
2
4
6
8
Mean of c
mean = 0.107
0.0
0.1
0.2
0.3
0.4
Mean of a
mean = 0.538
0.2
0.4
0.6
0.8
1.0
1.2
Mean of r
mean = 4.27
0
5
10
15
20
25
Mean of e
mean = 0.58
0.4
0.6
0.8
1.0
Mean of d

Figure 16

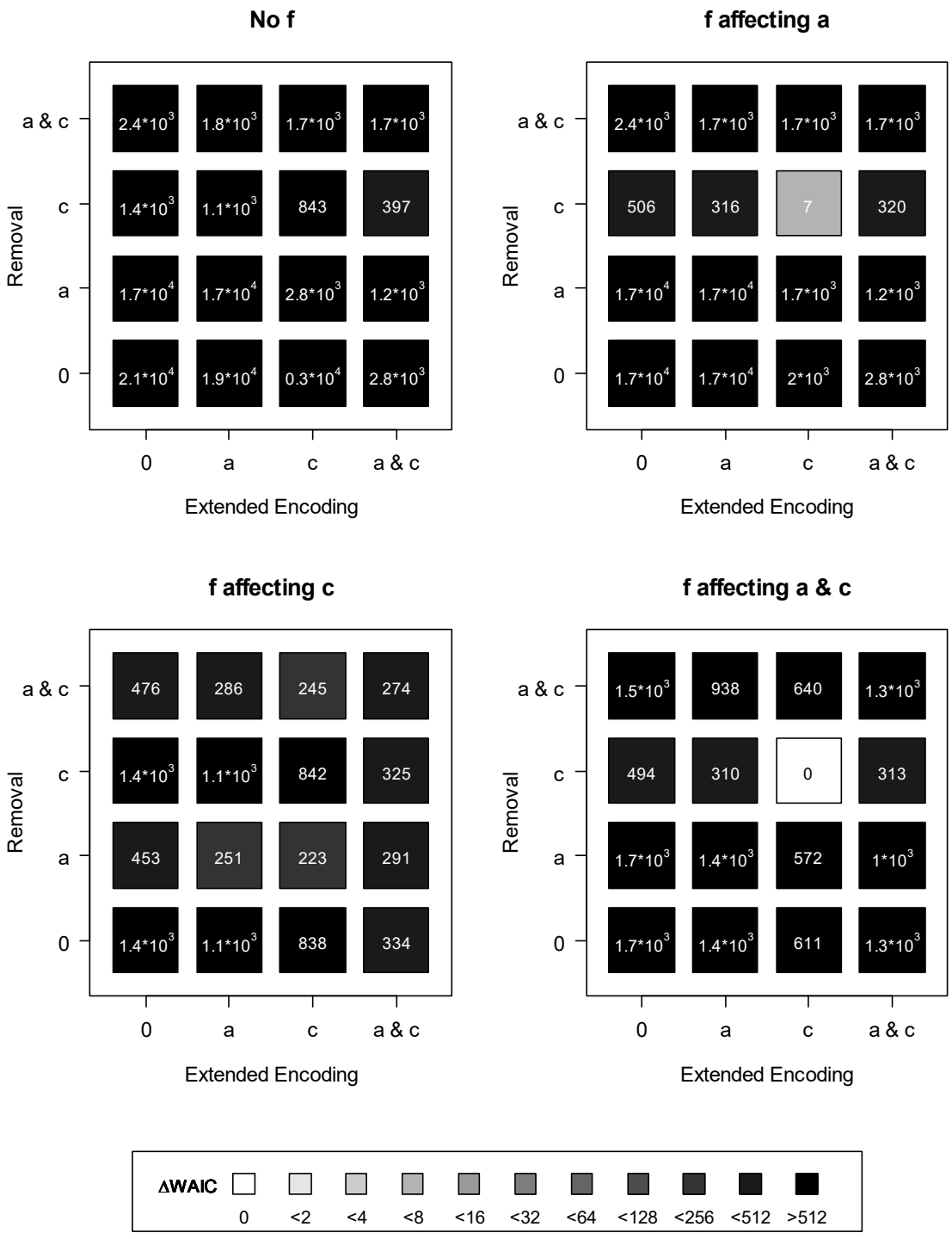

No f
f affecting a
f affecting c
f affecting a & c
Removal
Extended Encoding
0
a
c
a & c
2.4*10^3
1.8*10^3
1.7*10^3
1.4*10^3
1.1*10^3
843
397
1.7*10^4
2.8*10^3
1.2*10^3
2.1*10^4
1.9*10^4
0.3*10^4
2.8*10^3
506
316
7
320
2*10^3
476
286
245
274
842
325
453
251
223
291
838
334
1.5*10^3
938
640
1.3*10^3
494
310
313
572
1*10^3
611
ΔWAIC
<2
<4
<8
<16
<32
<64
<128
<256
<512
>512

Figure 17

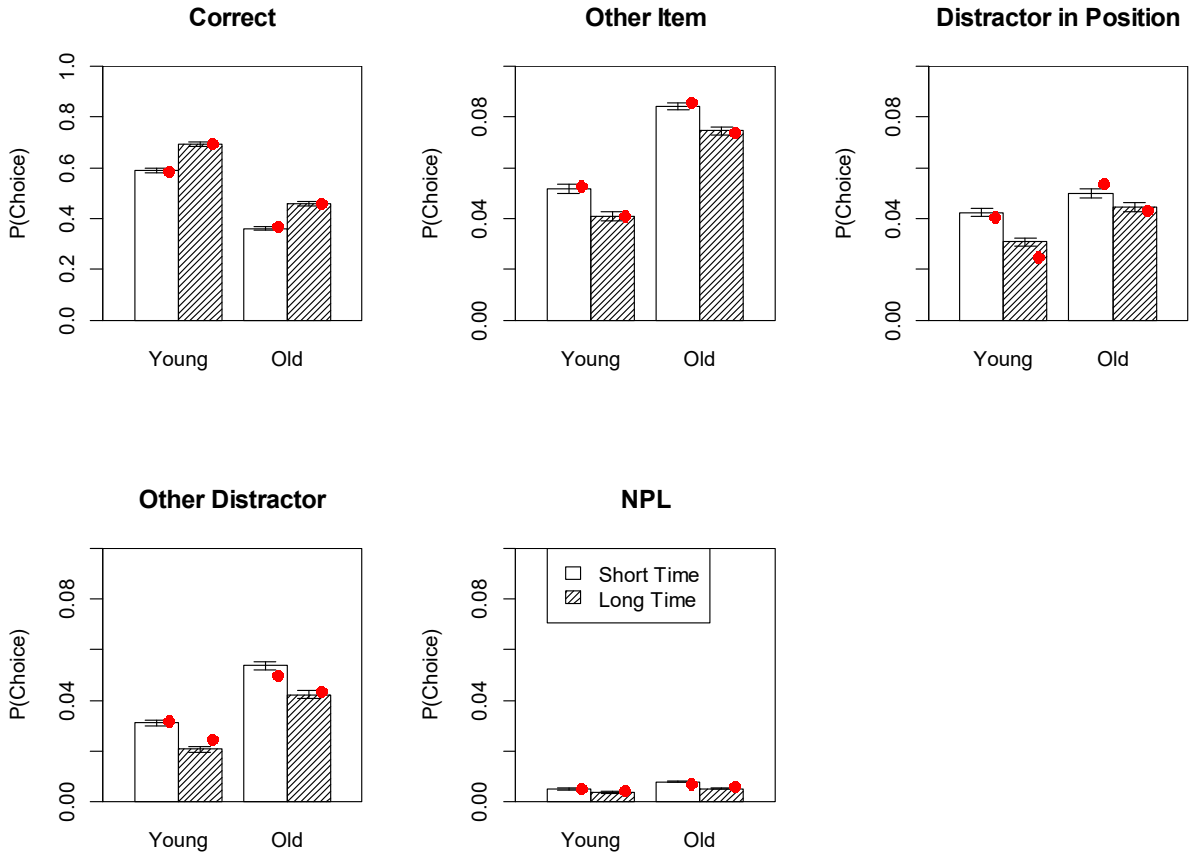

Correct
Other Item
Distractor in Position
Other Distractor
NPL
P(Choice)
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.04
0.08
Young
Old
Short Time
Long Time

Figure 18

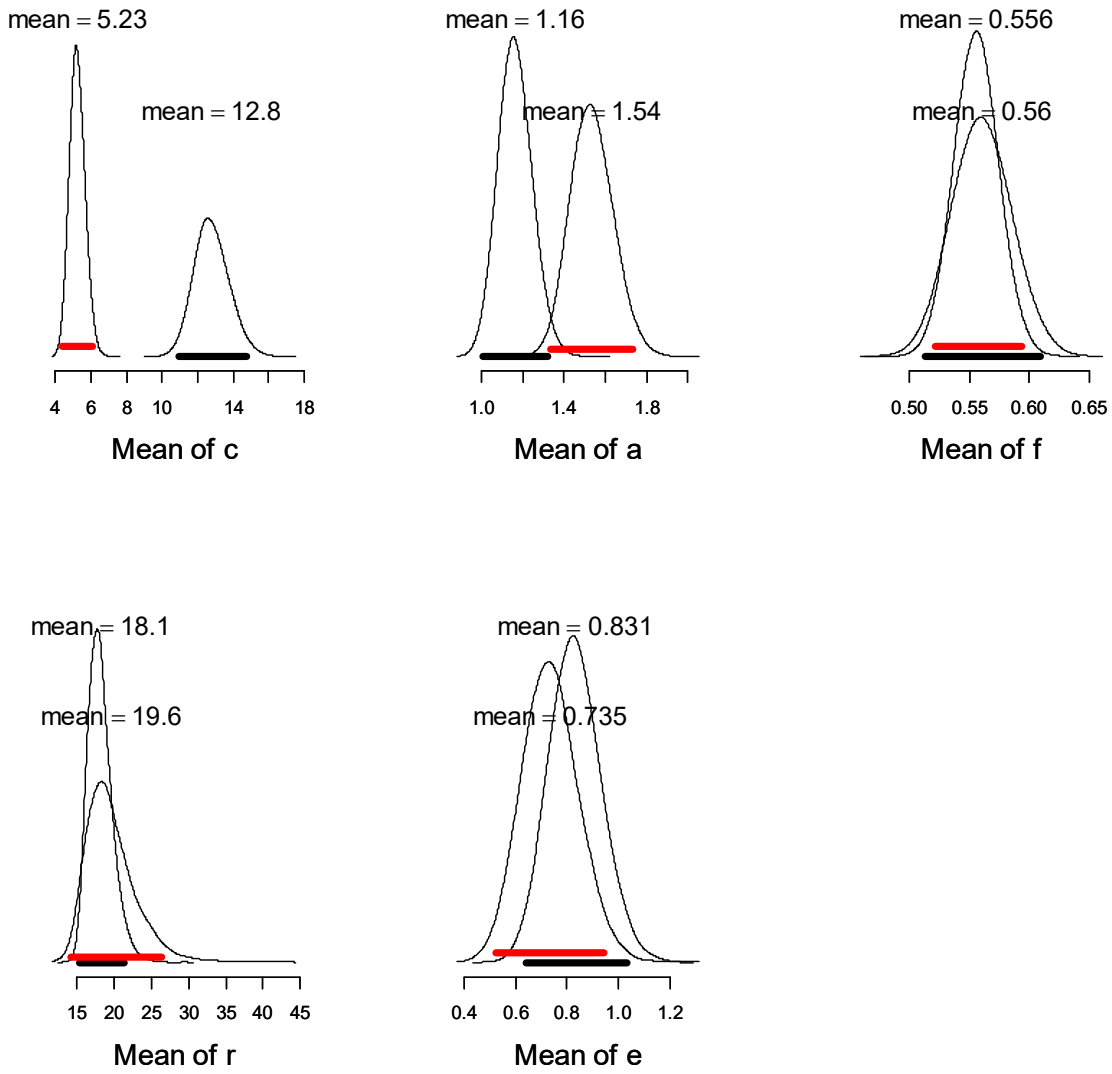

mean = 5.23
mean = 12.8
4 6 8 10 14 18
Mean of c
mean = 1.16
mean = 1.54
1.0 1.4 1.8
Mean of a
mean = 0.556
mean = 0.56
0.50 0.55 0.60 0.65
Mean of f
mean = 18.1
mean = 19.6
15 20 25 30 35 40 45
Mean of r
mean = 0.831
mean = 0.735
0.4 0.6 0.8 1.0 1.2
Mean of e

Figure 19

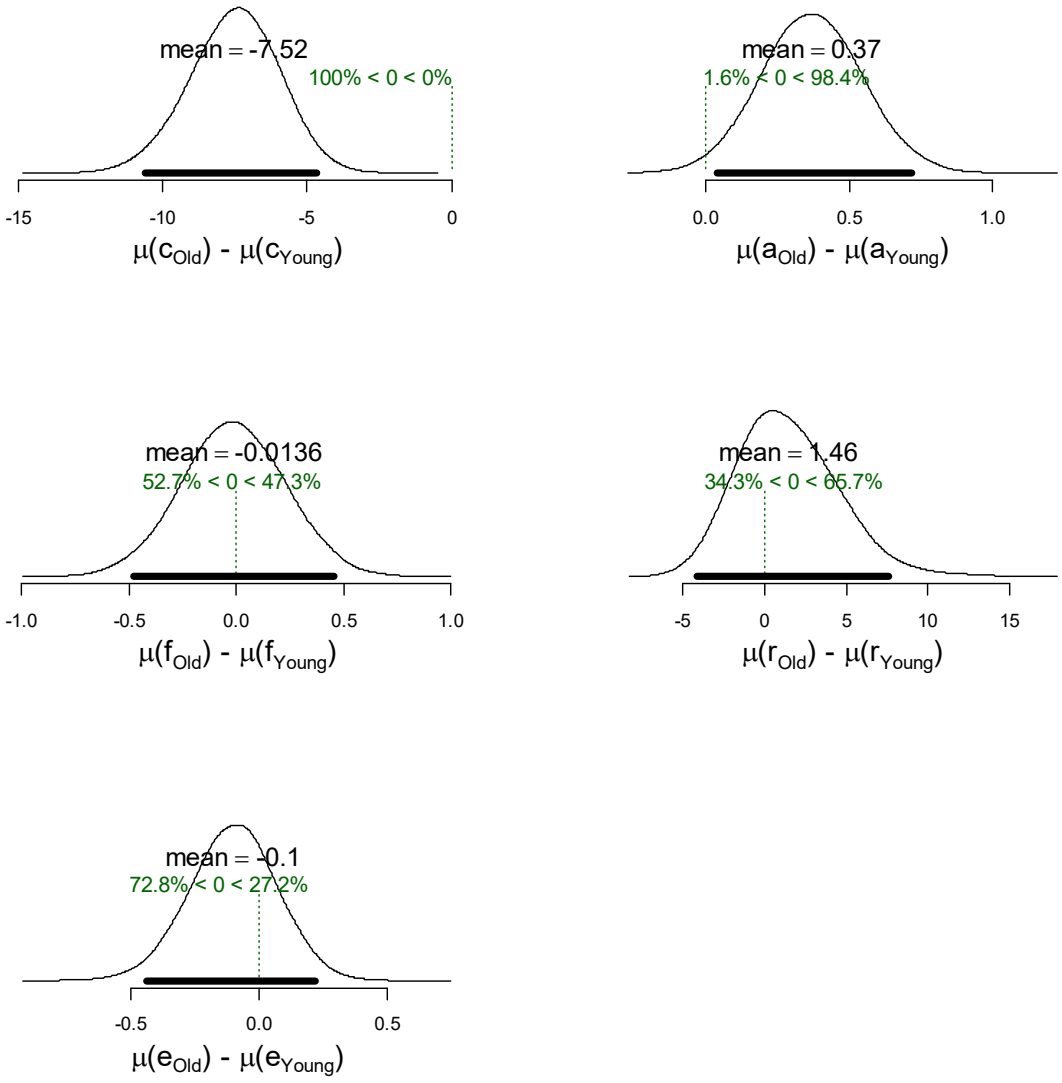

mean = -7.52
100% < 0 < 0%
-15
-10
-5
0
μ(c_Old) - μ(c_Young)
mean = 0.37
1.6% < 0 < 98.4%
0.0
0.5
1.0
μ(a_Old) - μ(a_Young)
mean = -0.0136
52.7% < 0 < 47.3%
-1.0
-0.5
0.0
0.5
1.0
μ(f_Old) - μ(f_Young)
mean = 1.46
34.3% < 0 < 65.7%
-5
0
5
10
15
μ(r_Old) - μ(r_Young)
mean = -0.1
72.8% < 0 < 27.2%
-0.5
0.0
0.5
μ(e_Old) - μ(e_Young)

Figure 20

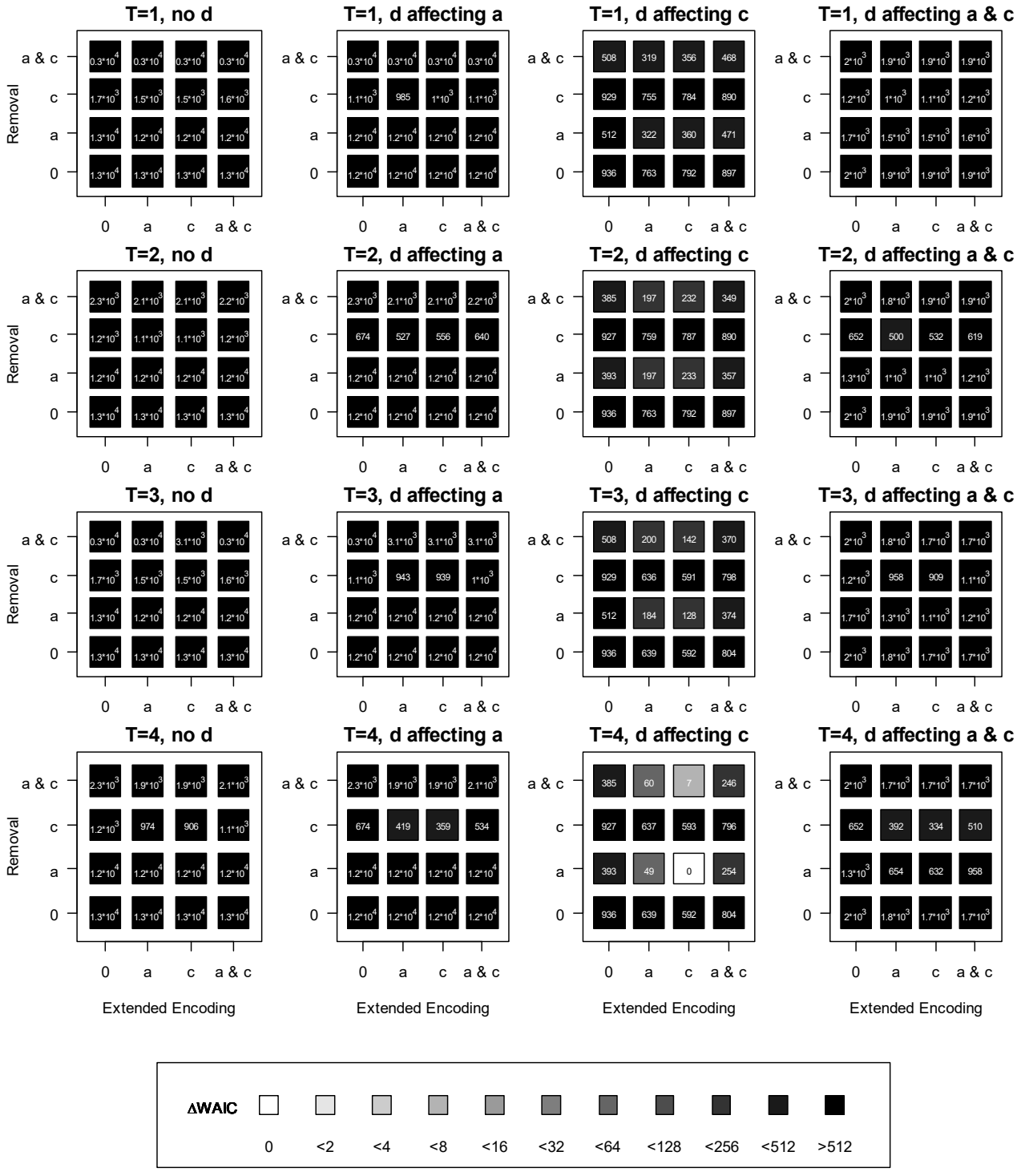

T=1, no d
T=1, d affecting a
T=1, d affecting c
T=1, d affecting a & c
T=2, no d
T=2, d affecting a
T=2, d affecting c
T=2, d affecting a & c
T=3, no d
T=3, d affecting a
T=3, d affecting c
T=3, d affecting a & c
T=4, no d
T=4, d affecting a
T=4, d affecting c
T=4, d affecting a & c
Removal
a & c
c
a
0
Extended Encoding
ΔWAIC
0
<2
<4
<8
<16
<32
<64
<128
<256
<512
>512

Figure 21

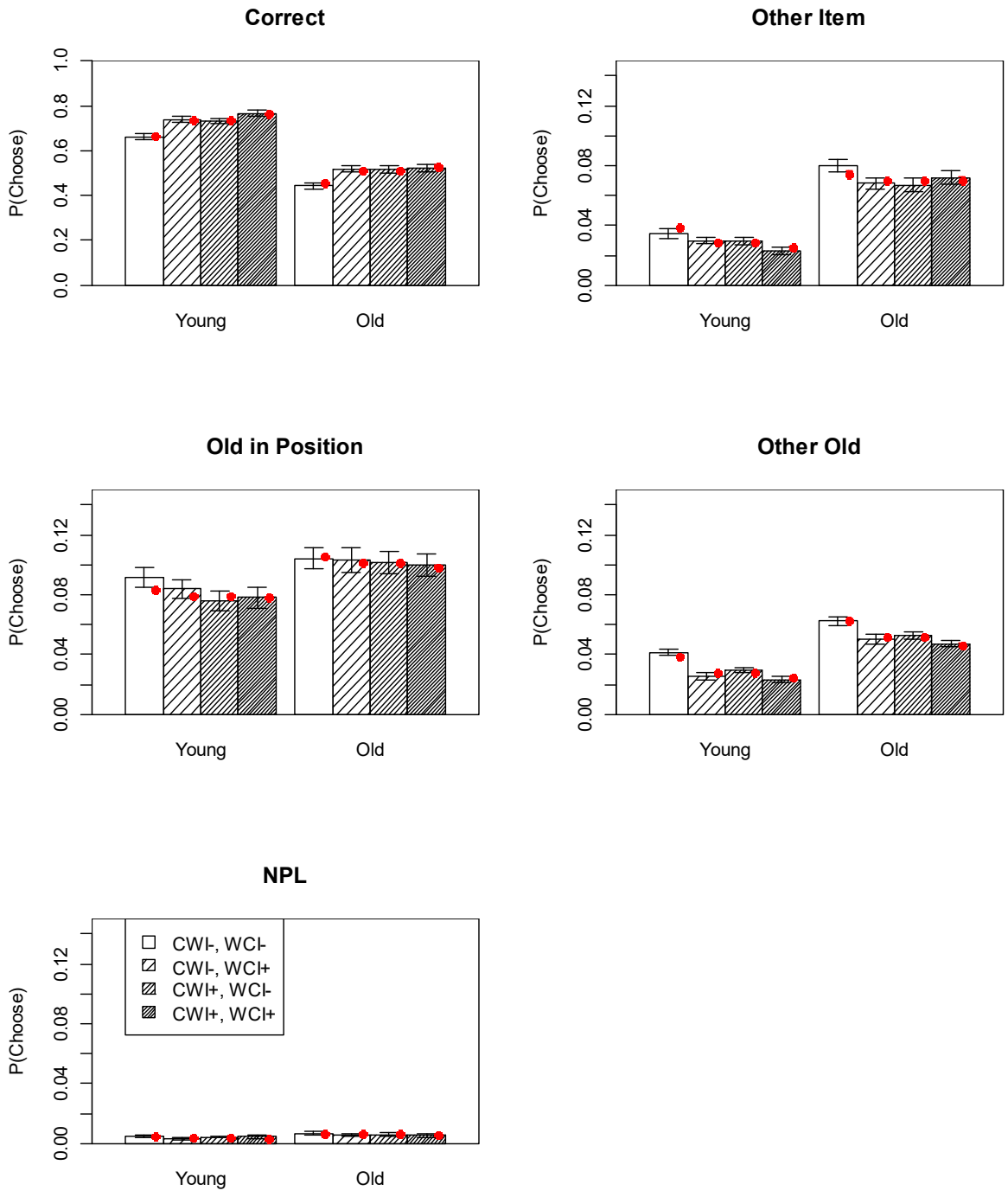

Correct
P(Choose)
0.0
0.2
0.4
0.6
0.8
1.0
Young
Old
Other Item
P(Choose)
0.00
0.04
0.08
0.12
Young
Old
Old in Position
P(Choose)
0.00
0.04
0.08
0.12
Young
Old
Other Old
P(Choose)
0.00
0.04
0.08
0.12
Young
Old
NPL
P(Choose)
0.00
0.04
0.08
0.12
CWI-, WCI-
CWI-, WCI+
CWI+, WCI-
CWI+, WCI+
Young
Old

Figure 22

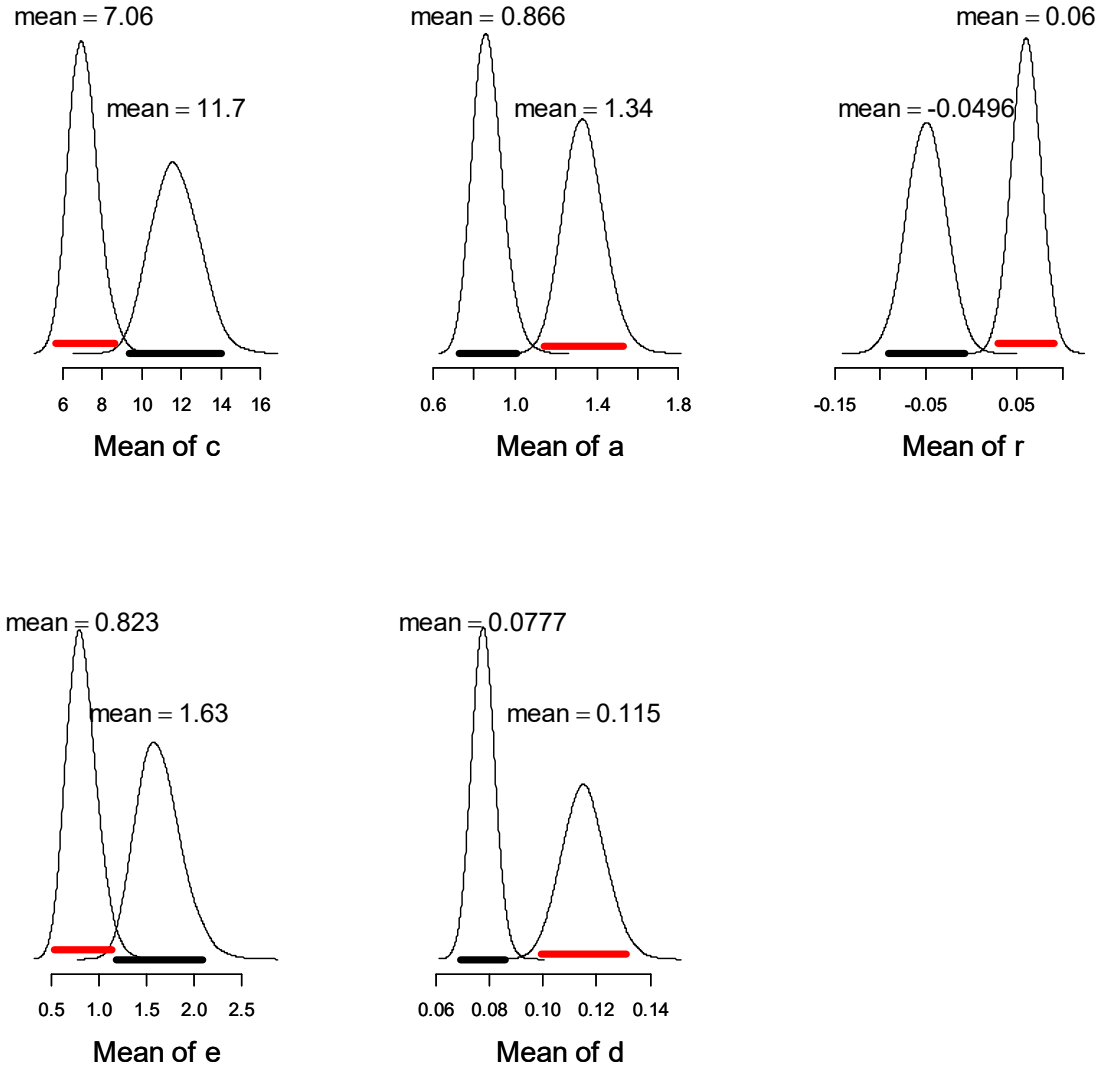

mean = 7.06
mean = 11.7
6 8 10 12 14 16
Mean of c
mean = 0.866
mean = 1.34
0.6 1.0 1.4 1.8
Mean of a
mean = 0.06
mean = -0.0496
-0.15 -0.05 0.05
Mean of r
mean = 0.823
mean = 1.63
0.5 1.0 1.5 2.0 2.5
Mean of e
mean = 0.0777
mean = 0.115
0.06 0.08 0.10 0.12 0.14
Mean of d

Figure 23

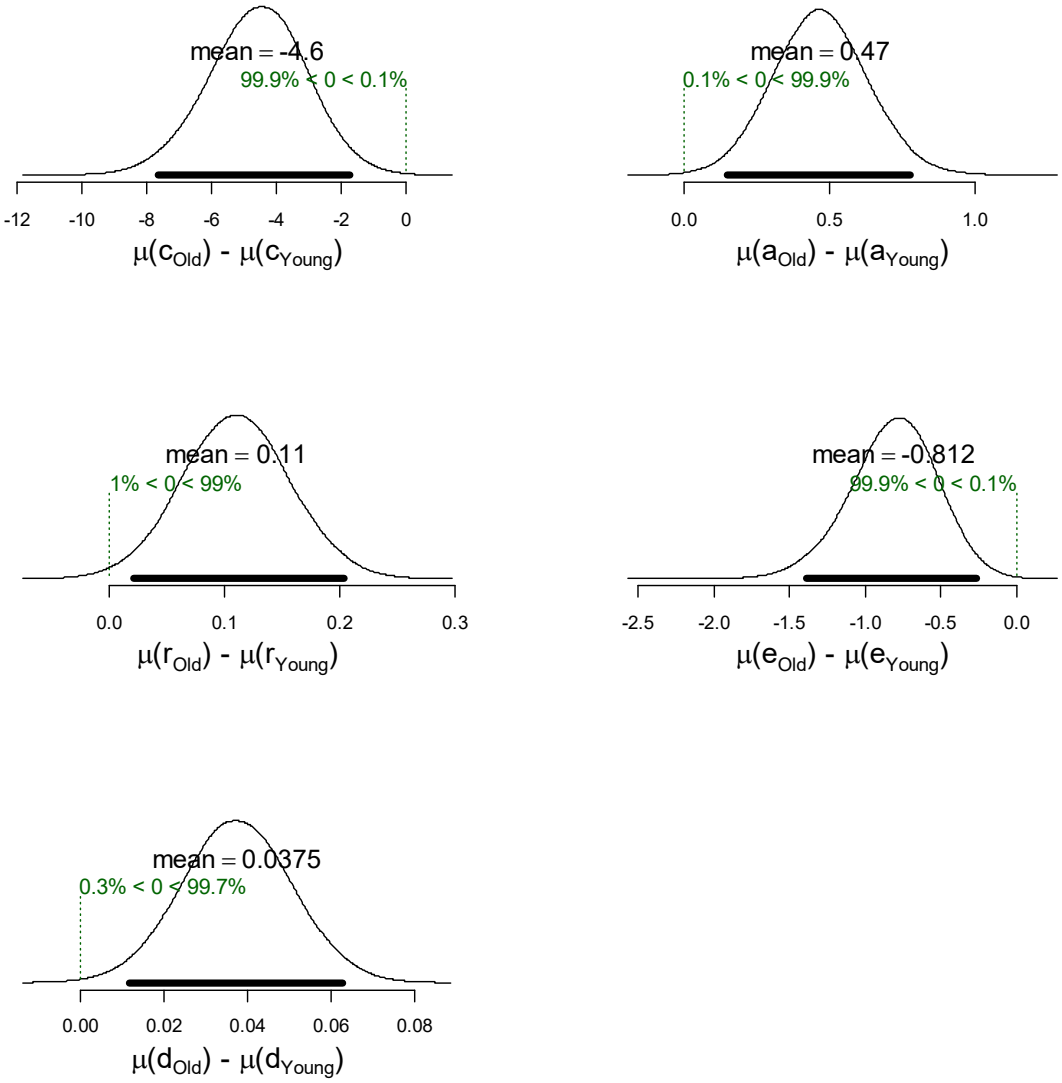

mean = -4.6
99.9% < 0 < 0.1%
-12 -10 -8 -6 -4 -2 0
μ(c_Old) - μ(c_Young)
mean = 0.47
0.1% < 0 < 99.9%
0.0 0.5 1.0
μ(a_Old) - μ(a_Young)
mean = 0.11
1% < 0 < 99%
0.0 0.1 0.2 0.3
μ(r_Old) - μ(r_Young)
mean = -0.812
99.9% < 0 < 0.1%
-2.5 -2.0 -1.5 -1.0 -0.5 0.0
μ(e_Old) - μ(e_Young)
mean = 0.0375
0.3% < 0 < 99.7%
0.00 0.02 0.04 0.06 0.08
μ(d_Old) - μ(d_Young)

Figure 24

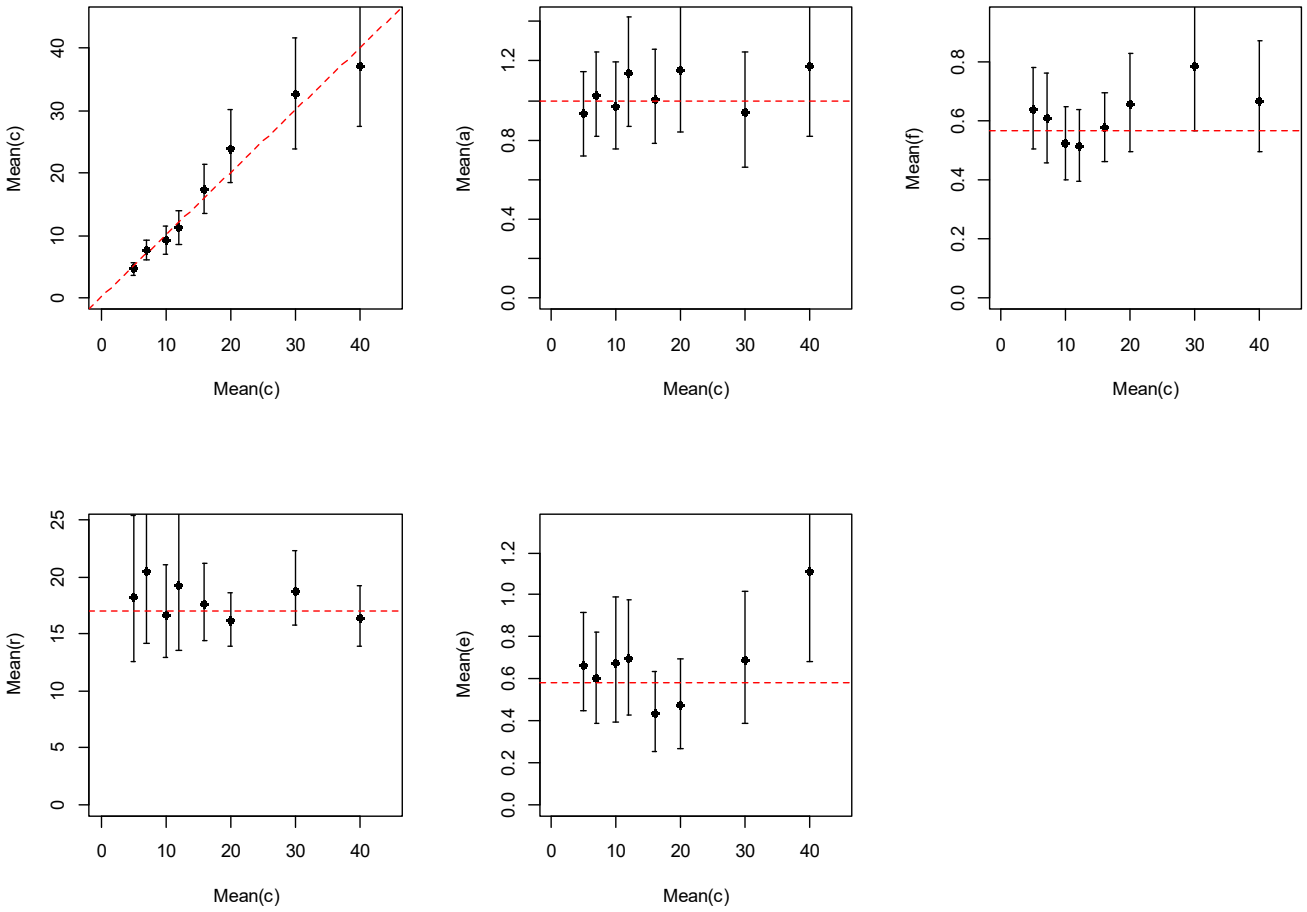

Mean(c)
Mean(c)
Mean(a)
Mean(c)
Mean(f)
Mean(c)
Mean(r)
Mean(c)
Mean(e)
Mean(c)

Figure 25

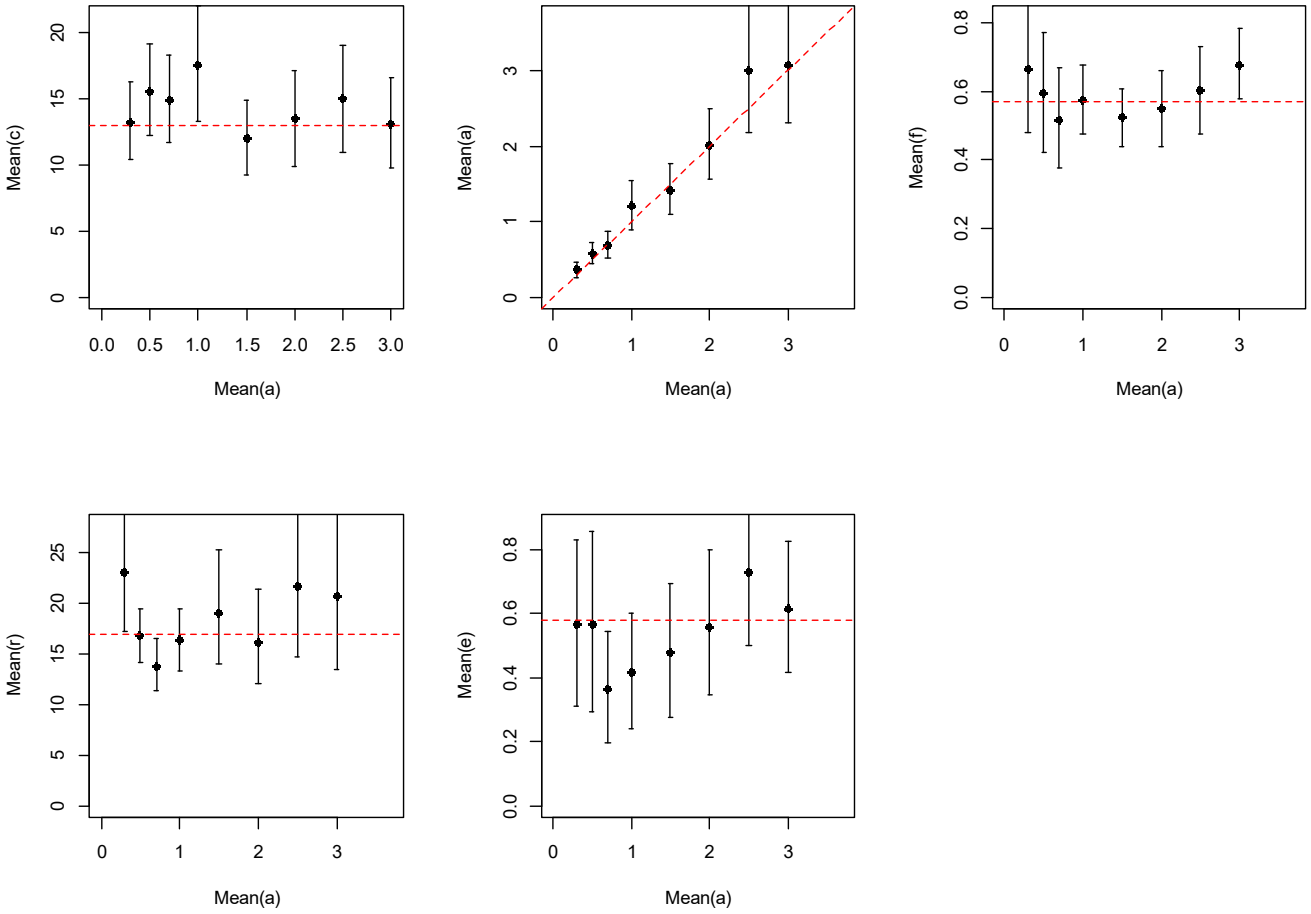

Mean(c)
Mean(a)
Mean(a)
Mean(a)
Mean(f)
Mean(a)
Mean(r)
Mean(a)
Mean(e)
Mean(a)

Figure 26

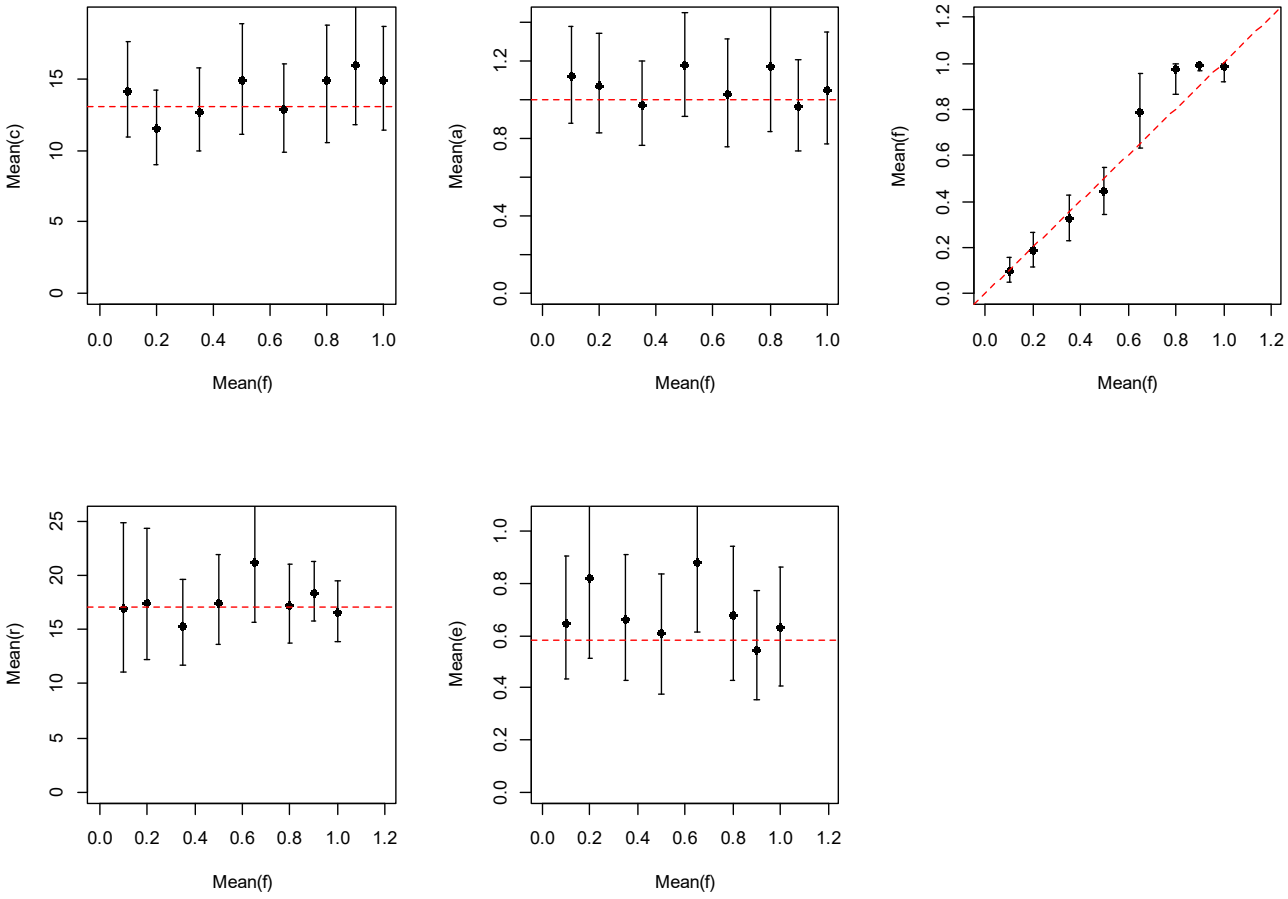

Mean(c)
Mean(f)
Mean(a)
Mean(f)
Mean(f)
Mean(f)
Mean(r)
Mean(f)
Mean(e)
Mean(f)

Figure 27

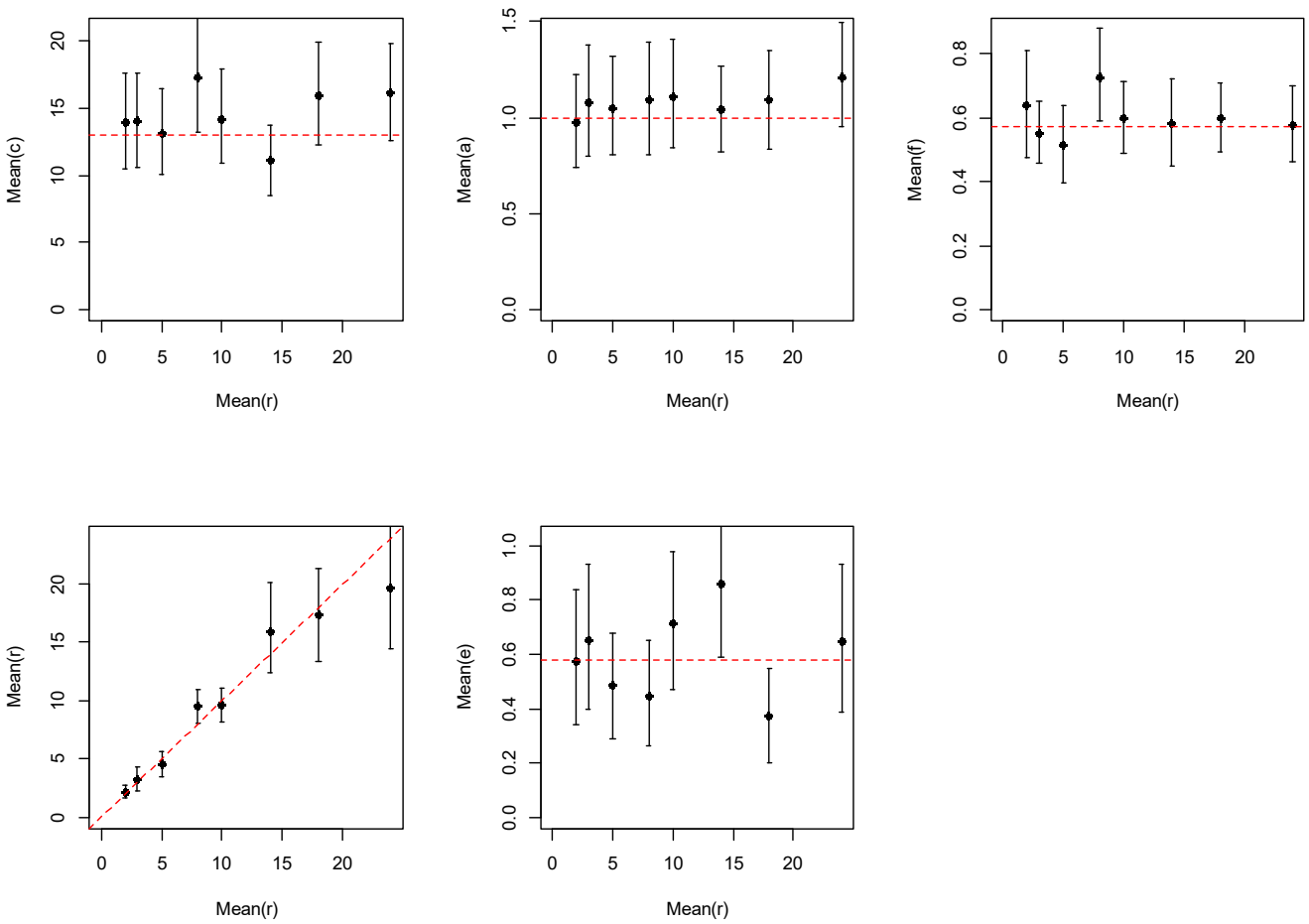

Mean(c)
Mean(r)
Mean(a)
Mean(r)
Mean(f)
Mean(r)
Mean(r)
Mean(r)
Mean(e)
Mean(r)

Figure 28

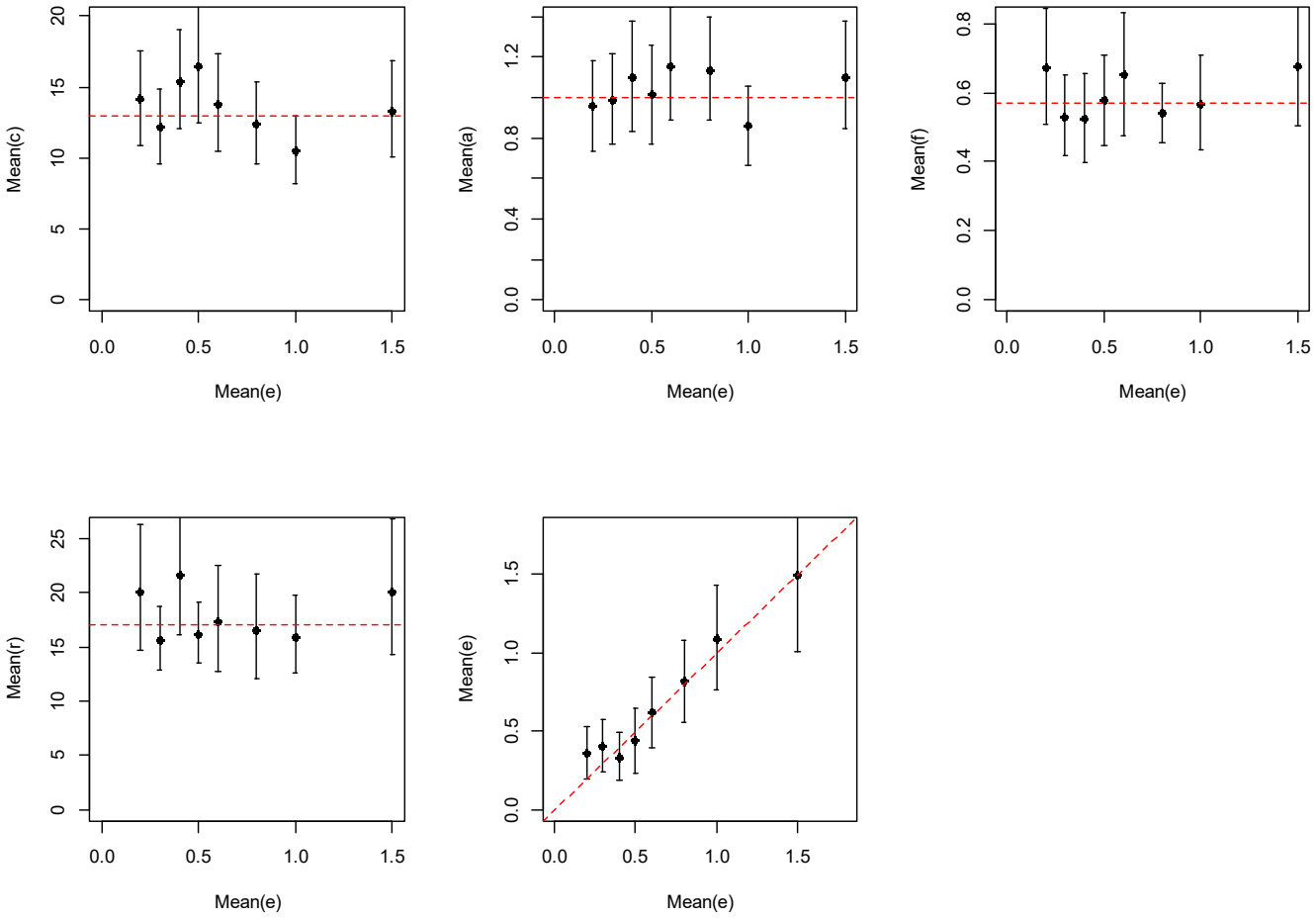

Mean(c)
Mean(e)
Mean(a)
Mean(e)
Mean(f)
Mean(e)
Mean(r)
Mean(e)
Mean(e)
Mean(e)

Figure 29

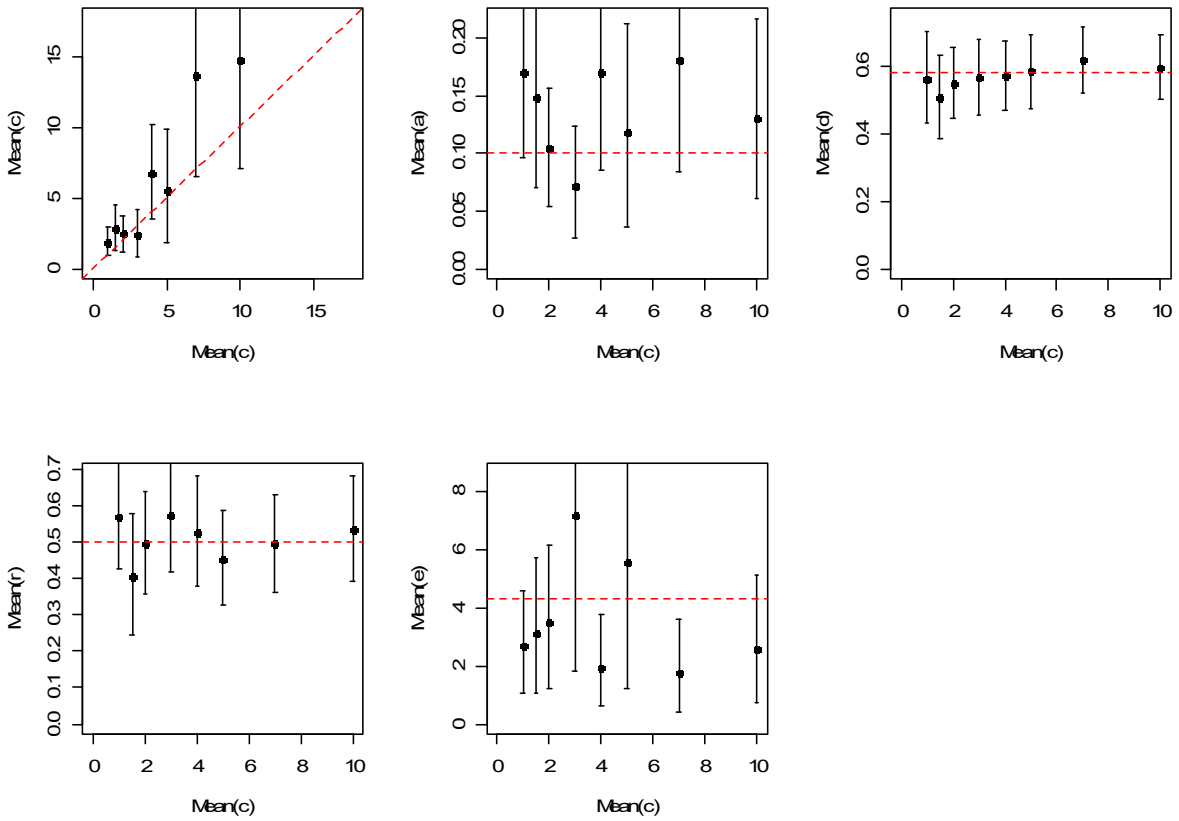

Mean(c)
Mean(c)
Mean(a)
Mean(c)
Mean(d)
Mean(c)
Mean(r)
Mean(c)
Mean(e)
Mean(c)

Figure 30

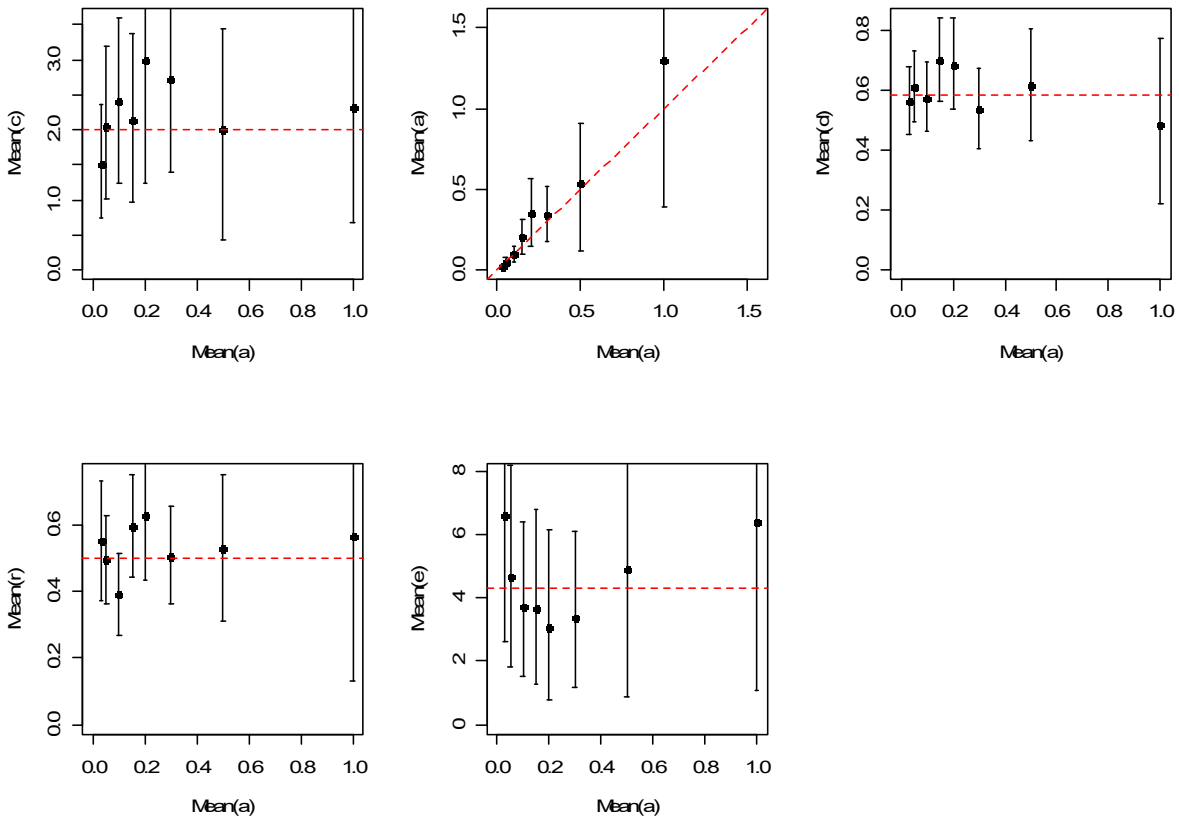


Figure 31

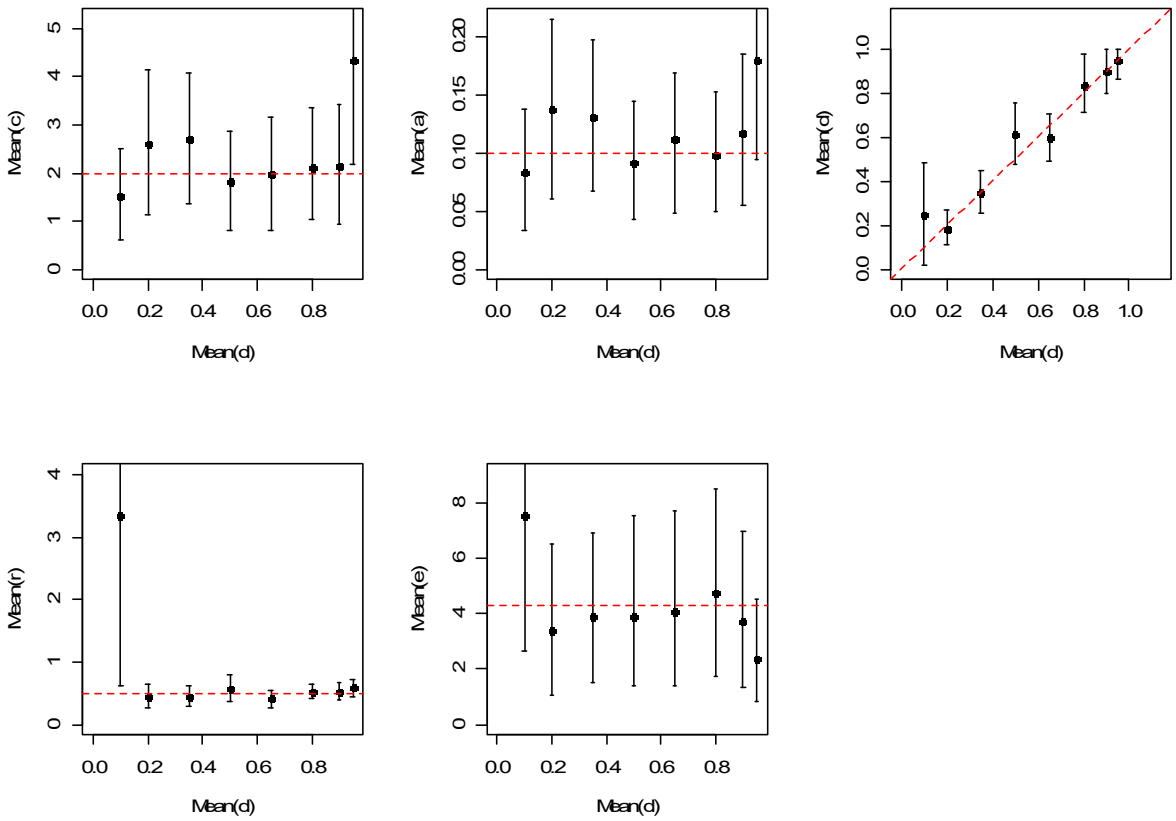

Mean(c)
Mean(d)
Mean(a)
Mean(d)
Mean(d)
Mean(d)
Mean(r)
Mean(d)
Mean(e)
Mean(d)

Figure 32

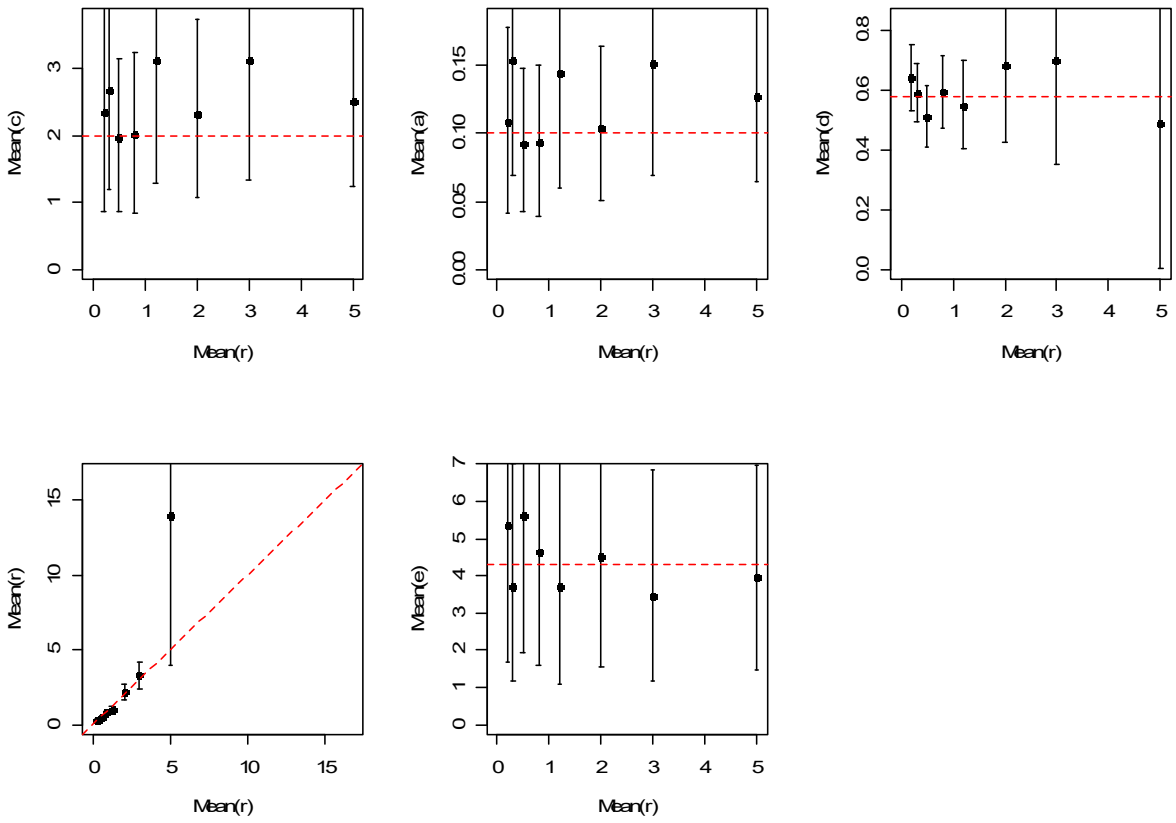

Mean(c)
Mean(r)
Mean(a)
Mean(r)
Mean(d)
Mean(r)
Mean(r)
Mean(r)
Mean(e)
Mean(r)

Figure 33

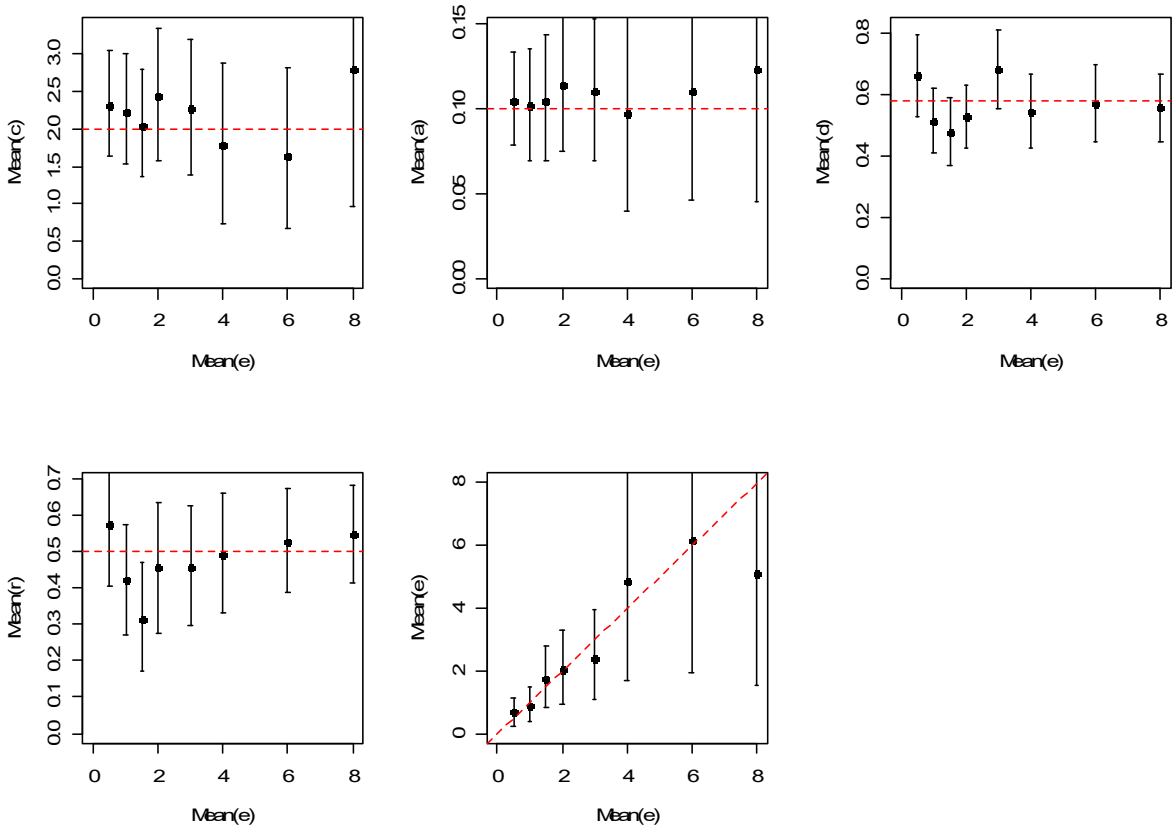

Mean(c)
Mean(e)
Mean(a)
Mean(e)
Mean(d)
Mean(e)
Mean(r)
Mean(e)
Mean(e)
Mean(e)

Figure A1

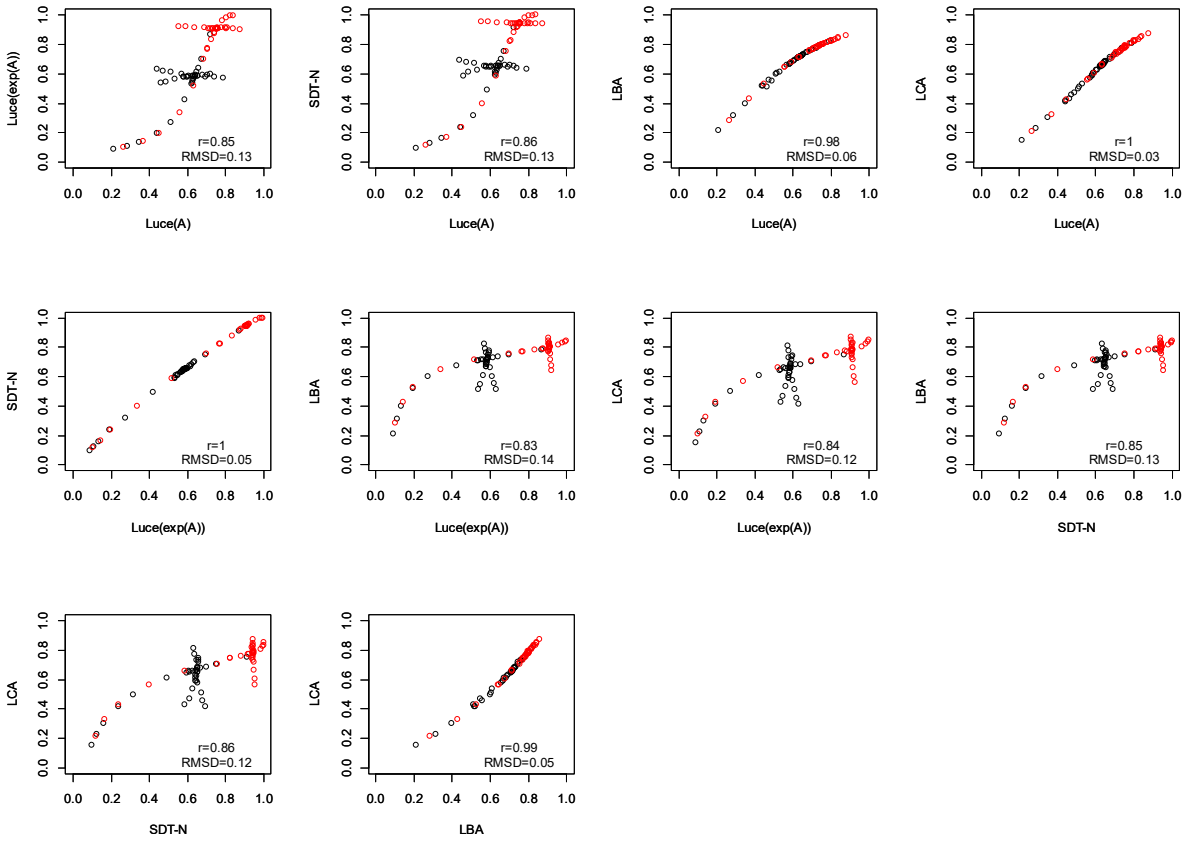
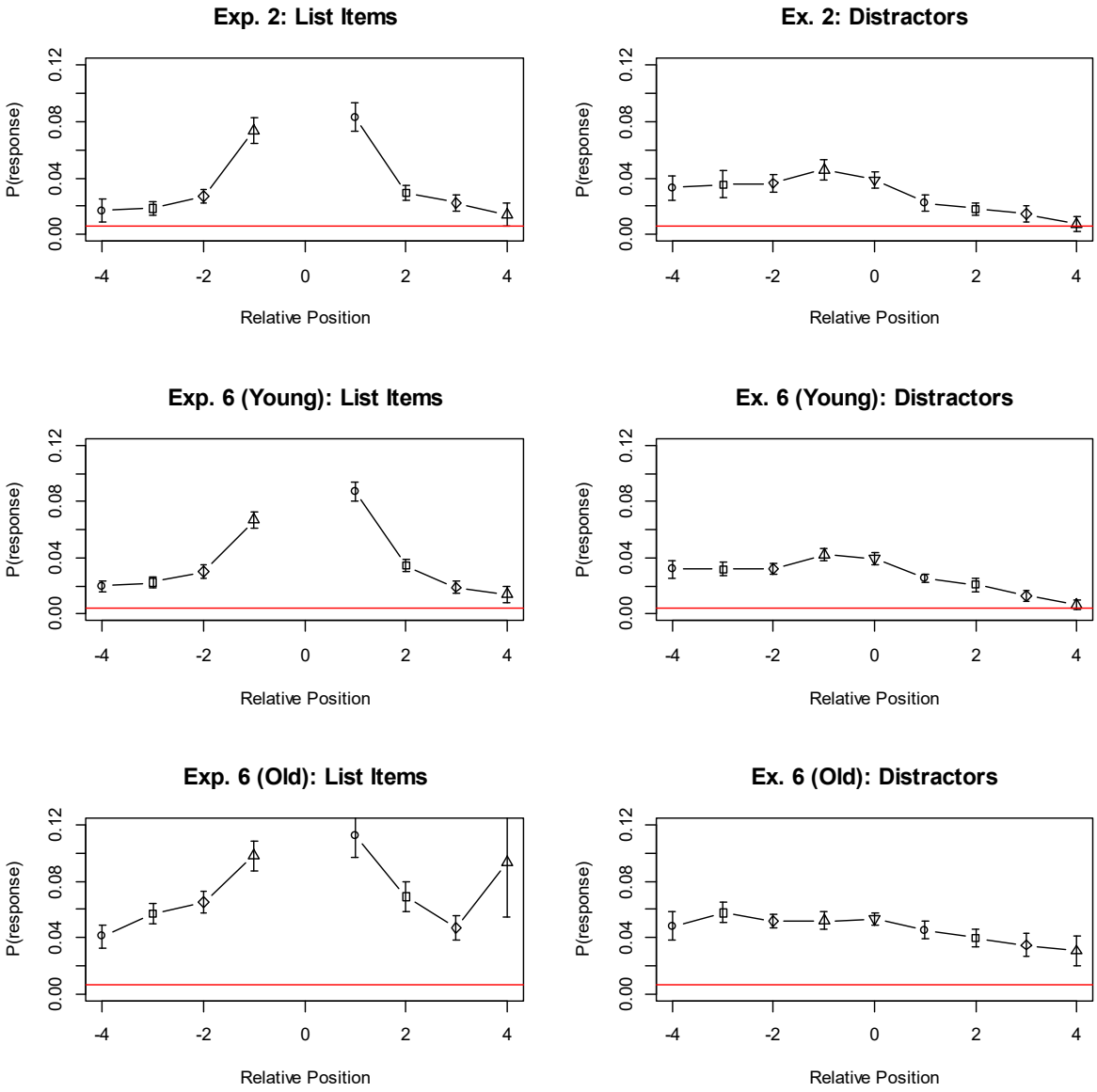

Luce(exp(A))
Luce(A)
r=0.85
RMSD=0.13
SDT-N
Luce(A)
r=0.86
RMSD=0.13
LBA
Luce(A)
r=0.98
RMSD=0.06
LCA
Luce(A)
r=1
RMSD=0.03
SDT-N
Luce(exp(A))
r=1
RMSD=0.05
LBA
Luce(exp(A))
r=0.83
RMSD=0.14
LCA
Luce(exp(A))
r=0.84
RMSD=0.12
LBA
SDT-N
r=0.85
RMSD=0.13
LCA
SDT-N
r=0.86
RMSD=0.12
LCA
LBA
r=0.99
RMSD=0.05

Figure A2


Exp. 2: List Items
Ex. 2: Distractors
Exp. 6 (Young): List Items
Ex. 6 (Young): Distractors
Exp. 6 (Old): List Items
Ex. 6 (Old): Distractors
P(response)
0.00
0.04
0.08
0.12
-4
-2
0
2
4
Relative Position